ESTIMATING ENGINEERING COST RISK USING LOGISTIC AND MULTIPLE REGRESSION

THESIS

Vincent P. Sipple, Captain, USAF

AFIT/GAQ/ENC/02-02

DEPARTMENT OF THE AIR FORCE
AIR UNIVERSITY

AIR FORCE INSTITUTE OF TECHNOLOGY

Wright-Patterson Air Force Base, Ohio

AFIT/GAQ/ENC/02-02

ESTIMATING ENGINEERING COST RISK USING LOGISTIC AND MULTIPLE REGRESSION

THESIS

Presented to the Faculty

Department of Mathematics and Statistics

Graduate School of Engineering and Management

Air Force Institute of Technology

Air University

Air Education and Training Command

In Partial Fulfillment of the Requirements for the

Degree of Master of Science in Acquisition Management

Vincent P. Sipple, BBA

Captain, USAF

March 2002

AFIT/GAQ/ENC/02-02

ESTIMATING ENGINEERING COST RISK USING LOGISTIC AND

MULTIPLE REGRESSION

Vincent P. Sipple, BBA
Captain, USAF

Approved:

//signed//	
Edward D. White, Major, USAF (Chairman)	date
//signed//	
William K. Stockman, Lt Col, USAF (Member)	date
//signed//	
Michael J. Seibel (Member)	date
//signed//	
Michael A. Greiner, Capt, USAF (Member)	date

Acknowledgements

I thank Maj Edward White, my thesis advisor, for all his effort and guidance in support of this research. His patience and direction have proven instrumental to the success of this effort. I couldn't have asked for a better advisor for my situation. I would also like to thank my committee members, Mr. Michael Seibel, Capt Michael Greiner, and Lt Col William Stockman. They all have given generously of their time to help in this effort.

I am also grateful to the many professionals who have given their assistance to include Lt Col Bradley Ayres, Mr. Roy LeBlanc, Mr. Kurt Bledsoe, Mr. Ken Teutsch, Mr. John Webb, Col William Nanry, Ms. Theresa Obrien, Ms. Sandy McCardle, Ms. Debbie Cann, Ms. Donna Snead, Mr. Harold Dagel, Dr. Dave Christensen, Capt Andrew Wallen, Ms. Janet Wentworth, Ms. Rosie Avant, Mr. Joe Escoto, Mr. Rick Mejia, and Mr. Bill Stranges.

I thank my fellow students for their friendship and help. I thank the US Air Force for allowing me to have the opportunity to be a student at AFIT; I thank my family for their endurance, understanding, prayers, and love; and I thank God for his providence in all aspects of this effort.

Vincent P. Sipple

Table of Contents

List of Figures

List of Tables

Abstract

This study explores a two-step procedure for assessing defense acquisition program cost growth using historical data. Specifically, we seek to predict whether a program will experience cost growth and, if applicable, how much costs will increase. We compile programmatic data from the Selected Acquisition Reports (SARs) between 1990 and 2000 for programs from all defense departments. We focus our analysis on cost growth in research and development dollars for the Engineering Manufacturing Development phase of acquisition. We further limit our study to only one of the seven SAR categories of cost growth – engineering cost growth. We explore the use of logistic regression in cost analysis to predict whether cost growth will occur. Using this methodology, we produce a statistically significant model that accurately predicts approximately 70 percent of our validation data. For those programs that have cost growth, we use a multiple regression model (an adjusted R^2 of 0.4645), with a natural log transformation, to predict the expected amount of cost growth. We discover the two-step logistic and multiple regression approach produces desirable results. Finally, we find schedule variables to have the most predictive ability from the 78 candidate independent variables analyzed.

ESTIMATING ENGINEERING COST RISK USING LOGISTIC AND MULTIPLE REGRESSION

I. Introduction

General Issue

The cost growth that major weapon systems incur throughout their acquisition life cycles concerns those who work in the acquisition environment. A 1993 study by RAND cites that by the time a system completes the production and fielding phase of acquisition, Department of Defense (DoD) Acquisition Category (ACAT) I programs historically experience an average cost growth of approximately 20 percent from initial estimates (Drezner, 1993:xiii).

Cost growth in major weapon system programs negatively impacts DoD, the country, and depending on the contract type, the DoD contractors involved. To successfully contain cost growth, program managers must carefully plan their program, coordinating with all stakeholders so that the plan developed encompasses all aspects of the user's needs. The more carefully considered and better coordinated the plan, arguably the less cost growth will occur. In support of this proposition, RAND notes that smaller DoD programs tend to actuate higher percentage cost growth than their larger counterparts; RAND cites as possible reason for this phenomenon the lower level of management scrutiny placed on smaller dollar value programs (Drezner, 1993:xii).

Aside from containing cost growth, DoD managers must also concern themselves with accurately identifying those risks related to potential cost increases in the program cost estimates. Managers can reduce *measured* cost growth by more accurately assigning dollar values to known risks, thereby increasing the accuracy of the baseline figure from which DoD measures cost growth. The cost estimating community supports management in this arena by doing its best to assign appropriate dollar amounts to the program-specific risk factors, then aggregating these dollar amounts into the cost estimate.

Specific Issue

Often, cost estimators use subjective means for assigning dollar amounts to risk factors, which they then use to incorporate estimated cost growth within the budget baseline estimate. Typically, a cost estimator solicits expert opinions on the overall risk levels of different aspects of a program and then uses a heuristic to apply dollar amounts to those risk values. A more objective method for assigning dollar values to risk factors involves a careful analysis of historical data. This approach requires a cost analyst to understand relationships between program attributes and observed cost growth. In such an approach, it might behoove the estimator to split cost growth into various categories to examine whether different types of cost growth have distinct sets of predictors. Statistical regression techniques prove useful in determining such relationships, and this research applies such techniques to find predictors of cost growth.

Scope and Limitations of the Study

The Selected Acquisition Reports (SARs) are a collection of individual program reports that (among other things) capture all of the cost variances on many major defense acquisition programs. These reports provide an adequate data source from which to analyze cost growth. Due to both the accessibility and the detail of the SARs, we use them to build a database for our research. The SARs separate program cost variance into seven categories: Economic, Quantity, Estimating, Engineering, Schedule, Support, and Other (Drezner, 1993:7). The demarcation of these seven components allows for a standardized comparison of variances across programs, and a more meticulous analysis of cost growth. The SARs also contain a variety of other programmatic details that lend to their usefulness in a detailed analysis of cost growth. In general, these details include major schedule milestone dates, physical and performance characteristics, and contractual information. As with other databases, our SAR database has limitations, but none that preclude its use for this research.

In this study, we measure cost growth as a percentage increase in cost from the Development Estimate (DE) as recorded in the SAR format. We limit our study to cost growth in the Research and Development, Test and Evaluation (RDT&E) accounts during the Engineering and Manufacturing Development (EMD) phase of acquisition. We further scope our effort to only consider one of the seven categories of cost variances as delineated in the SAR reports - cost variances due to engineering changes. This category includes cost growth that occurs as a result of physical changes in the end item (Knoche, 2001:22; Drezner, 1993:7). Thus, we only explain one piece of the cost-growth puzzle, but prepare the way for potential completion of this puzzle with our compilation of information on all other categories of cost growth within our database and through our validation of methodologies.

For reasons of time constraints and of data currency, the study includes only programs that use the DE as the baseline estimate and programs whose Engineering Manufacturing Development (EMD) phase of acquisition falls within the period 1990-2000. We only use one SAR per program and choose the most recent available. In many cases, the most recent DE-based SAR available is the last SAR of the EMD phase of acquisition. Quirks exist in the SAR data that further limit the research (e.g. security classification, etc.) Chapter III addresses many of these limitations in depth. Finally, the DE may already include some unknown budget for risk, which limits the interpretation of the results of this research.

Past research looks at cost growth within the DoD from a macro perspective. High-level decision-makers use these studies for macro-level reasons, such as finding general trends in overall cost growth. As such, much of the past research has a descriptive rather than an inferential statistical focus. Though ours is an inferential study, we use these historical studies to help us find candidate predictor variables for cost growth. We find only a few historical studies that apply multiple regression, and none consider logistic regression techniques. Our study explores several new frontiers as we investigate a newly created database, compile an extensive list of candidate predictor variables derived from past research, forge a unique approach at analysis with both logistic and multiple regression, and address cost growth in a new way - at the constituent level.

Research Objectives

This study has three main objectives. First, the study explores the utility of logistic regression in finding predictors of engineering cost growth. To our knowledge, no researcher has explored the use of logistic regression in cost analysis before. Specifically, we use logistic regression to determine if certain program characteristics predict whether a program experiences engineering cost growth in the RDT&E budget during the EMD phase of development. Logistic regression differs from multiple regression in that it predicts a binary response. In our case the binary response is: *Does a program experience cost growth, Yes or No?* Second, the study seeks to find predictors of the degree to which cost growth occurs. We use multiple regression to determine if certain program characteristics predict the amount of engineering cost growth in the RDT&E budget in the EMD phase of development. Lastly, we seek to discover the nature of these predictive relationships such that one may use the formulas to predict whether a program will have cost growth and to predict point and range estimates of the percent of engineering cost growth in the RDT&E budget in the EMD phase of program development.

Chapter Summary

This study attempts to leverage off past cost growth research to create models that meet the needs of the financial management community to better estimate risk in dollar terms according to program characteristics. To develop these models, we perform logistic and multiple regressions on data from programs recorded in the SARs over the past decade. The study involves only engineering cost growth in the RDT&E budget as measured from the DE of the program. While managers must deal with cost growth in many ways, this study seeks to reduce measured cost growth by helping cost estimators more accurately estimate cost growth early in the program.

II. Literature Review

Chapter Overview

This chapter provides an overview of the research involving cost growth. We first describe the overall acquisition, cost-estimating, and risk assessment environment, then follow with details of previous studies that relate to the topic of the study at hand. From the information gathered in this chapter, we develop a historical and logical framework from which to begin building predictive regression models.

The Acquisition Environment

Peter Woodward, in his thesis, mentions factors in the acquisition program management environment that may cause a program to overrun. Woodward talks of the constrictive nature of the DoD acquisition funding environment, "For example, it is impossible to take advantage of quantity buys and other cost-saving techniques when program managers are required to obligate all their funds within a year or two of their appropriation" (Woodward, 1983:106). Woodward alludes to funding rules that require the obligation (putting funds on a contract) of research and development funds within a period of two years and the obligation of procurement funds within a period of three years. Woodward further states, "It is also difficult to obtain these cost savings when a manager does not even know for certain whether his program funding will be cut from one year to the next" (Woodward, 1983:106). Program managers that fail to obligate funds within these time windows face dangers that range from chastisement and program restructure to loss of funding and program cancellation. In fact, even when a program manager manages his program perfectly, upper management might choose to sacrifice his program in order to bail out another higher priority program that has funding problems. One can see that in this environment, as program schedules slip, program managers face increasing pressure to sacrifice cost-effectiveness for expediency.

In a cost growth study, one must consider the diversity of programs that exist within the acquisition environment. The *DoD Manual Cost Analysis Guidance and Procedures* lists the following as categories of Defense Acquisition Systems: "Aircraft, Engines, Missiles, Ships, Tanks and Trucks, Data Automation/ADPE, and Electronics." The manual further divides electronics into the following four subcategories: "Radar, Communications, Satellite, EW [electronic warfare]" (Department of Defense, 1992:13-14). Then, the manual details types of "key system characteristics and performance parameters" that prove useful in estimating each particular category of acquisition (Department of Defense, 1992:13-14). This categorization hints that heterogeneity characterizes DoD acquisition cost estimating such that different types of systems have different drivers of cost behavior. Cost growth as measured from the Development Estimate for different categories of acquisition systems may also have this heterogeneous property.

The Cost Estimating Environment

Cost growth has proven a significant problem for some time for program offices. During the early eighties, the Reagan administration recognizes two ways to control the problem of cost growth. "Despite some initial steps, controlling cost growth remains a major problem. The solution must include more realistic estimates accurately reflecting future costs and difficult choices to reduce requirements when costs grow" (Office of the Under Secretary of Defense, 1981:4). According to this quote, creating estimates that are more realistic provides one way of controlling cost growth. Cost/requirements tradeoffs provide a second way to control cost growth. The second method has since come into vogue through the "cost as an independent variable" or CAIV approach to program management (Ayres, 2000:3). In CAIV, cost takes on greater importance when making programmatic decisions. This more than likely has led to successful cost control, although quantifying that success proves elusive. This research seeks to enhance the first method of cost control, "more realistic estimates."

High-level DoD management personnel continue to concern themselves with cost growth. In December of 2000, Air Force experts brief the Chief of Staff of the Air Force on their findings in a focused cost growth study of 16 current ACAT I programs. The study observes cost growth (as recorded in the SARs) that occurs over the years 1997-1999 and ignores cost decreases. This study finds that cost growth from quantity and schedule changes accounts for 32 percent and 24 percent of the cost growth (respectively) in these 16 programs (Westgate, 2000:3). Estimating changes account for 20 percent of the growth; engineering changes account for 17 percent of the growth; and changes in support costs account for seven percent of the total cost growth. This study shows that the overwhelming majority of cost growth results from budget decisions and requirements changes. These decisions and changes come from Air Force Headquarters, DoD Headquarters, or Congress. In many of these programs, programmatic problems seem to instigate these budget decisions and requirement changes (Westgate, 2000:6).

Cost growth for the 16 programs over the three-year period totals 12 percent ($20.2 billion) (Westgate, 2000:14).

The researchers offer several recommendations to help control cost growth. To gain better visibility into the "cost of delay" that occurs when production rates change due to quantity or schedule changes, the researchers suggest including those costs in the quantity or schedule variances in the SAR (Westgate, 2000:16). The visibility of the cost of making quantity or schedule changes will help decision-makers avoid such changes when the cost exceeds the supposed benefits.

The researchers recognize that high cost growth results not only from poor visibility of cause-effect relationships, but also in the limiting of decision-makers' options. Along this vane, the researchers recommend that headquarters "limit fenced modernization dollars to preclude funding instability" (Westgate, 2000: 16). 'Fencing' (i.e. restricting the use of dollars) minimizes flexibility, and decision-makers tend to make poorer funding decisions in this inflexible funding environment, because fewer options exist which they can pursue.

As a third recommendation, the researchers suggest that the Air Force "require highest priority projects to be estimated and funded at a higher confidence level" (Westgate, 2000: 16). This suggestion alludes to the practice of calculating and quantifying cost risk within the weapon system cost estimate used to produce the budget profile. In this process, estimators use a probability distribution to determine the probability of occurrence and impact of events that might increase cost. These researchers advocate using some level of confidence above 50 percent to ensure that the top priority programs receive the funding they need without having to 'rob' from other programs. This may prove especially useful for high dollar programs, where a small percent increase can mean a drastic evaporation of funding.

The researchers espouse a fourth recommendation that invites criticism: the rewarding of program managers for cost performance (Westgate, 2000: 16). From strictly a cost perspective this incentive makes sense, but some would argue that without proper care in instituting such a policy, the result could be an imbalanced priority on cost above performance and schedule to such a degree as to jeopardize the delivery of a product that the war fighter needs within the appropriate timeframe.

Another suggestion by the team, that the Air Force "optimize program schedules instead of subjecting to budget constraints," faces great resistance by program managers under the current politics of the acquisition-funding environment (Westgate, 2000:17). Once Congress approves a funding profile, many program managers would rather hold on to what money they have in the years they have the money than risk trading their money for money in a different year in order to gain possible cost savings. This results from the uncertainty inherent in the funding environment stemming from stories of program managers who gave up funding in one year expecting a return of funds in the next, but who failed to receive the promised funds.

Finally, the research team recommends to the Chief of Staff that he "create an integrated system to capture standard budget, execution and performance data across [the] AF Modernization Program" (Westgate, 2000: 17). This recommendation reiterates the need to better capture, standardize, and disseminate information to make smarter decisions that should result in minimized cost growth. In summary, this study identifies cost growth as a problem visible to the highest levels of Air Force leadership, and it identifies possible ways to better control the problem.

Risk and Uncertainty in Cost Estimating

Documenting Uncertainty in Estimates

The Office of the Secretary of Defense (OSD) Cost Analysis Improvement Group (CAIG) gives guidelines for documenting cost estimating uncertainty for DoD system acquisition programs. First, they mandate that "areas of cost estimating uncertainty will be identified and quantified" (Department of Defense, 1992:22). Programs must document this uncertainty in the Cost Analysis Requirements Document (CARD). Second, the CAIG prescribes "the use of probability distributions or ranges of cost" to quantify uncertainty (Department of Defense, 1992:22). Third, they ask that the uncertainty estimated be "attributable to estimating errors" (Department of Defense, 1992:22). They give the following examples:

> …uncertainty inherent with estimating costs based on assumed values of independent variables outside data base ranges, and uncertainty attributed to other factors, such as performance and weight characteristics, new technology, manufacturing initiatives, inventory objectives, schedules, and financial condition of the contractor… (Department of Defense, 1992:22)

In addition to uncertainty, the DoD procedures also provide for the estimation of contingencies and sensitivity analysis. For contingencies, the manual gives the estimator the option to include a contingency amount or to exclude such an amount. If the estimator includes an amount for contingencies, he must give the reason for the contingency estimate as well as the rationale for the estimate. In addition, he must "include an assessment of the likelihood that the circumstances requiring the contingency will occur" (Department of Defense, 1992:22). This of course, implies the association of a probability distribution with such circumstances.

The Nature of Risk Analysis

Within the cost estimating community, differing opinions exist as to the meanings of risk and uncertainty. Rather than attempting to champion one definition over another, this paragraph seeks only to substantiate a particular distinction between risk and uncertainty to serve as a common starting point for discussion of risk analysis in this paper. Webster's defines risk as "the possibility of loss or injury," and defines 'uncertainty' as "the quality or state of being uncertain." To avoid defining a word with a form of itself, one must again search the dictionary to find that 'uncertain' means "not certain to occur," or "not known beyond doubt." Thus, from these definitions, one can infer that both risk and uncertainty share within their meanings the idea of 'questionable occurrence'. However, the definition of 'risk' adds to that 'questionable occurrence' the aspect of 'harm' through the words, "loss or injury." Thus, for the purpose of this paper, 'risk' involves both 'questionable occurrence' and 'harm,' while 'uncertainty' simply embodies 'questionable occurrence' within its definition.

The DoD cost estimating community considers cost growth as the "increase in cost of a system from inception to completion," and it considers cost risk as "the funds set aside to cover predicted cost growth" (Coleman, 2000:3). Thus, the cost risk represents the projected dollar amounts associated with risk, while the cost growth represents the incurred dollar amounts associated with the risk (Coleman, 2000:3).

The *AFMC Financial Management Handbook* gives the Air Force perspective on risk analysis:

> Cost estimating deals with uncertainty. What the analyst attempts to do is to describe in the best terms possible the probability distribution of the cost event in the future. One value for the cost estimate is the result of one prediction of that future event. Risk Analysis is a careful consideration of the areas of uncertainty associated with future events. The preferred common denominator for translating risk identified in the program is dollars. The detailed analysis of the risk to the program leads to better information for Air Force and other Government decision makers. (*AFMC Financial Management Handbook*, 2001:11-12)

Thus, risk analysis addresses the range of possible outcomes and their probabilities. The handbook distinguishes program risk as "the uncertainties and consequences of future events that may affect a program" (*AFMC Financial Management Handbook*, 2001:11-12).

The *AFMC Financial Management Handbook* recognizes three parameters for risk: technical, schedule, and cost risk. The handbook suggests that the estimator estimate the risk in these areas in terms of dollars and establish a probability distribution for each area. The program manager must decide from these distributions which number to use as the most appropriate number to add as part of the final cost estimate. All services use similar procedures, such that each service uses some logical method to assess risk in different areas of a program and quantify that risk within their estimates.

The handbook mentions three methods for handling risk analysis: a posteriori, a priori, and subjective judgment:

1) The first method, a posteriori, or "after the fact" relationship to past events (direct knowledge), is based on some previous occurrence such as the cost outcome of previous projects conducted by the organization. If enough samples from the past history (the population) are drawn, the probability of the next event occurring in a particular way may be estimated. A complex methodology like Monte Carlo simulation may also be used. The Monte Carlo simulation is conducted where the analyst determines the probability of future events by using an experimental model to

approximate expected actual conditions. Such a model is fashioned from previous histories of similar projects.

2) Sometimes a distribution of possible outcomes for an event is not based on experience or sampling but on a priori, or "before the fact" theoretical probability distribution. The use of the closeness of the assumptions used in developing the theoretical distribution is to the real world situation being analyzed.
3) Many times an analyst will have to use a subjective judgment (indirect knowledge) in estimating probability. This approach relies on the experience and judgment of one or more people to create the estimated probability distribution. The result is known as a subjective probability. A distribution estimate is an analysis by one or more informed persons of the relative likelihood of particular outcomes of an event occurring. Distribution estimates are subjective. An example of this approach is the Delphi method. (*AFMC Financial Management Handbook*, 2001:11-12)

Cost estimates in the SAR database may already include some dollar amounts within their budgets for risk. Program offices generally include amounts for risk within their budget submissions; however, higher-level reviews frequently result in removal of risk dollars from estimates.

Risk Assessment Methods

Several methods of risk assessment exist in the military cost estimating community. Use of different methods depends on the type of risk estimated, the level of detail needed in the estimate, the accuracy needed in the estimate, the timeframe within which the estimator has to complete the estimate, the skill of the estimator, the data and tools available to the estimator, and any office policies directing estimating practices.

In 1993, the RAND Corporation produces a study on DoD acquisition program cost growth. This study reiterates the need for cost risk-estimation techniques. In this study, the researchers make the following statement: "Unfortunately, no proven method exists to identify overly optimistic or pessimistic cost estimates at the different stages of a development program" (Drezner, 1993: 1). The RAND researchers also state, "Both overruns and underruns reduce the quality of resource allocation decisions" (Drezner, 1993: 1). Thus, the challenge exists in the form of creating a method for program offices to model cost overruns and underruns and incorporate such amounts in cost estimates.

Although both cost overruns and underruns adversely affect successful program management, the RAND study shows that estimates are systematically biased low. Therefore, systems managers face the dangers of cost overruns more often than the danger of cost underruns. The authors of the study point out the dangers of a downward bias in cost estimating:

> Systematic bias can lead to erratic acquisition decisions (e.g., more start and continuation decisions) that contribute to problems later in the system life cycle, such as the "bow wave" phenomena in which too many programs reach high funding levels at the same time: reduction in operation and support accounts to compensate for increases in the development and procurement accounts and quantity reductions that affect force structure plans and capabilities. (Drezner, 1993: 2)

<u>Carlucci Initiatives Recognize the Need for a Method</u>

In 1981, Deputy Secretary of Defense Frank C. Carlucci implements the Carlucci initiatives that seek to reform DoD program management. Among the 31 initiatives, two seek to improve the budgeting function by directing cost estimators to account for technological and other risk factors in their estimates (Woodward, 1983:6). To implement this policy, the Office of the Secretary of Defense encourages the use of the Total Risk Assessing Cost Estimate (TRACE) methodology as a possible means for incorporating risk into estimates to account for possible cost growth (Office of the Under Secretary of Defense, 1981:11-1). Peter Woodward in his thesis addressing funds management in the face of risk and uncertainty, appropriately characterizes the challenge of creating an estimate for program risk (management reserve):

> Thus, the hidden issue concerning management reserve is not the use of such a reserve itself, but the perception that Congress and higher-level management have of a service's program to accurately manage risk and uncertainty without exceeding the budget constraints as defined in the program. In order to achieve this, more objective statistical techniques can be used to derive the baseline cost estimate. Therefore, the present system of submitting a point cost estimate (which includes fixed program milestones, fixed schedule, and fixed performance parameters) must be modified so that additional information gained as the program progresses can be used to get the necessary funds for its completion. At present, once a point estimate is submitted, it becomes the controlling guideline throughout the life of the program. (Woodward, 1983:105)

In this passage, Woodward recognizes the power of statistical techniques to achieve objective estimates for management reserves. Woodward also recognizes that such a technique must have the flexibility to apply to different stages in the program life cycle. A technique that can achieve this objectivity and flexibility, Woodward claims, would not only produce more accurate estimates, but also give Congress and the Executive Branch more confidence in the program office's ability to manage its funding properly, presumably making management reserves less susceptible to retraction.

Types of Risk Methods

Figure 1 shows the different risk assessment techniques recognized by the Ballistic Missile Defense Organization (BMDO) cost estimating community. The chart shows how the "degree of precision" needed in an estimate drives the type of estimate used: as the degree of precision needed increases, the estimate techniques used become more detailed and difficult (Coleman, 2000:4).

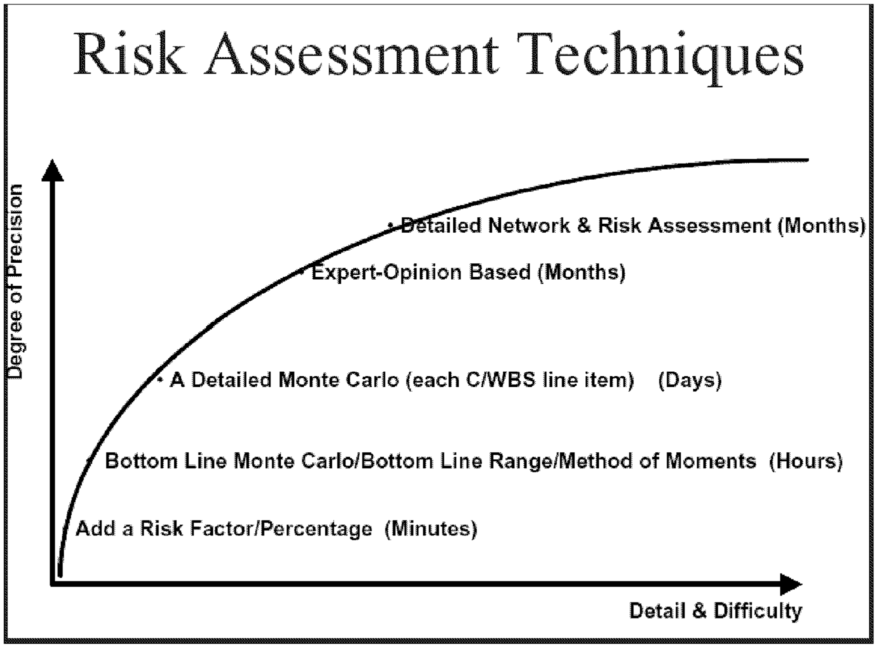


Figure 1. Risk Assessment Techniques (Coleman, 2000:4-9)

Starting from the most difficult and most precise end of the spectrum, the Detailed Network and Risk Assessment technique requires a very detailed schedule and task breakout. This method assigns either beta or triangular distributions to the schedule item durations to create a stochastic model from which to estimate the risk of a schedule slip. The estimator uses the Monte Carlo Simulation method to estimate the cost (Coleman, 2000:4-9).

The Expert-Opinion-Based technique represents the next level of detail down from the network technique. This method relies on surveys of experts to determine the possible distributions of Work Breakdown Structure (WBS) item costs. This method also uses Monte Carlo simulation to estimate a range of possible costs. It relies on the abilities of the experts to accurately assess the situation in light of their past experiences;

"the problem is whether technical experts have any real sense of how much things cost, or how much costs can rise" (Coleman, 2000:12).

The technique of the next difficulty level down is the Detailed Monte Carlo Simulation for "each C/WBS line item," where C/WBS is the Cost or Work Breakdown Structure of the program (Coleman, 2000:4). Although the previous two methods use Monte Carlo Simulation, this method differs from the previous two in that it relies on historical databases of cost and other programmatic information from which to develop probability distributions of cost outcomes (Coleman, 2000:16). This method quickens the process by avoiding lengthy surveys or PERT analyses, but its weakness lies in the applicability and currency of the data used in the database (Coleman, 2000:17). Despite these weaknesses, this method gives a reasonable amount of accuracy for the amount of time that an estimator puts into it, as Figure 1 depicts (Coleman, 2000:4).

The Bottom Line Monte Carlo, Bottom Line Range, and Method of Moments techniques in Figure 1 represent estimating on a less detailed level (Coleman, 2000:4). These methods may use Monte Carlo Simulation, but on higher levels of the WBS. These methods might use a limited database or analogy methodology to determine risk estimates, or they might use expert opinion to determine risk estimates. The least precise and easiest technique, "Add a Risk Factor/Percentage," relies on technical expert judgment to assign a high-level, subjective risk factor for the estimate (Coleman, 2000:4).

Monte Carlo Simulation

Arguably, the most favored method for estimating uncertainty, Monte Carlo Simulation provides a capability to the cost estimator that adds rigor to subjective estimates. Monte Carlo software exists that ties the probability distributions for multiple programmatic cost risk items to those items within the cost estimate. The software requires that the estimator define the probability distribution for each of the risk items, and gives the estimator considerable amount of flexibility in terms of the choice of probability distributions. The estimator can enter the parameters of a probability distribution based on either subjective judgments or a historical database. Once the estimator specifies the distributions for each risk item, the software runs the Monte Carlo Simulation. This simulation randomly generates results for each risk item specified, consistent with the assigned probability distributions. The software combines the results to display the overall program cost risk. This process repeats for a user-determined number of iterations, such that an overall cost risk distribution results. In such a fashion, the estimator finds a point estimate and a range of possibilities with their associated probabilities of occurrence (Coleman, 2000:5).

Past Research in Cost Growth

Before analyzing the data, we consider logical relationships in the program management environment that might explain cost growth. Past research helps in the search for explanations for cost growth. In this section, we describe various studies that address cost growth.

<u>RAND Study (2001)</u>

In a study in support of the Joint Strike Fighter program, RAND studies the effect of competition on the amount of cost growth that occurs in both the RDT&E and procurement budgets (Birkler, 2001:74). The researchers analyze 14 programs that use competitive strategies and 44 programs that do not use competitive strategies (Birkler, 2001:74). They find that "the results are mixed and the differences between the competitive and noncompetitive development [and procurement] CGFs (cost growth factors) are not statistically significant at the 10-percent level" (Birkler, 2001:80). Although it might prove enlightening to explore competitive programs versus non-competetive programs in a multiple regression study of the cost growth associated with engineering changes, we do not pursue that course of analysis in this study, largely due to unavailability of the required data.

BMDO Study

A recent BMDO cost growth study provides insight into the nature of cost growth. Using an internal BMDO database of programs (created from a subset of the SAR database), BMDO finds that RDT&E cost growth averages 21 percent while that of production averages 19 percent (Coleman, 2000:19). The study also shows that from seven to 16 percent of programs complete at or below the target cost (see Figure 2) (Coleman, 2000:19). From Figure 2, it appears at first glance that the lower the dollar value of a program, the greater the likelihood of a large cost growth factor. The author does not provide any statistical tests to explore this possibility, but the graph at least does not provide evidence against the idea.

The researchers of BMDO compare their cost growth results with past studies using the SAR database, evidencing a general commonality of cost growth factors (see Table 1) (Coleman, 2000:20). Differences in the results possibly stem from differences in the subsets of the SAR data used and differences in the methods used (Coleman, 2000:20). The study shows evidence of bias in cost risk estimates as described in the following sentences.

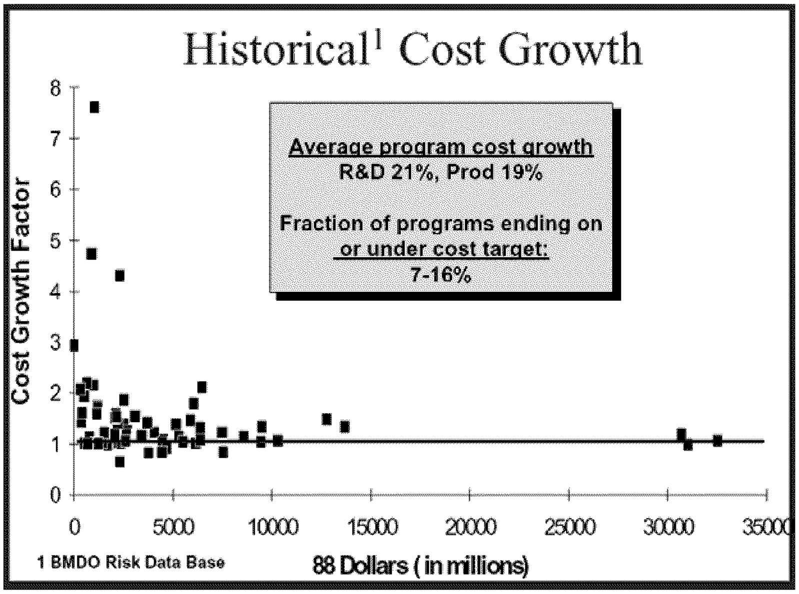


Figure 2. Historical Cost Growth (Coleman, 2000:19)

As a program progresses, cost estimators revise their estimates to reflect realized values of risk. The estimators reduce the amount of risk estimated and increase the cost estimate in other areas to reflect this change. Under the assumption of unbiased risk estimates, one would expect that realized risk would tend to equal the estimated risk on average given a large sample. In fact, the study shows that the risk portion of the estimate decreases at a slower rate than the rate of the rest of the estimate increases (Coleman, 2000:22-23). This evidences a general trend of underestimating risk.

This study does not break down cost growth into its components. In addition, this study does not distinguish cost growth by acquisition phase. Thus, we cannot specifically tie the results of the BMDO study to the nature of engineering cost growth in EMD; however, the study does give us general insight into predictors to pursue.

Table 1. Historical Cost Growth (Coleman, 2000:20)

Source	Raw Average			$ Wtd Average			N	During Prod
	Tot	R&D	Prod	Tot	R&D	Prod		
RAND 93	1.30			1.20	1.25	1.180	100+	1.02
CAIG 91	1.33	1.40	1.25	1.21	1.24	0.119	27	
TASC 94		1.49	1.54				20+	
TASC 96		1.43	1.55		1.21	1.350	14	0.99
Christensen 99				1.09	1.14			1.06

NAVAIR Study

NAVAIR presents its most recent study on cost growth at the 2001 DoD Cost Analysis Symposium, corroborating some of the results of previous studies, and adding new insight into cost growth. Their study assesses cost growth as reported in the SARs. As part of their analysis, they explore the possible need for "cohort tracking" when analyzing cost growth (Dameron, 2001:7). Webster's Collegiate Dictionary defines "cohort" as "band or group." By "cohort tracking," the NAVAIR team refers to the grouping of cost growth according to certain programmatic characteristics that relate to common patterns of cost growth. The team divides program cost growth into five categories or cohorts – RDT&E cost growth for programs with a planning estimate (PE) and a development estimate (DE); RDT&E cost growth for programs with a DE only; procurement cost growth for programs with a PE, a DE, and a production estimate (PdE); procurement cost growth for programs with a DE and a PdE only; and procurement cost growth for programs with a DE only (Dameron, 2001:10).

Cost estimators perform each of the three different possible estimates (PE, DE, and PdE) at a different phase in the acquisition life cycle. The estimator performs a PE for a Milestone (MS) I review, the first action in the Program Definition Risk Reduction (PDRR) phase. The estimator performs a DE for a MS II review, the first event in the EMD phase of the acquisition life cycle. Finally, the estimator may or may not perform a PdE (sometimes the DE suffices) for a MS III review, the first event in the procurement phase of the acquisition life cycle. Not all programs use all three of the above-mentioned program phases, and one discerns the program structure from the types of estimates used. The NAVAIR team does not explicitly state, but we presume that they use the five cohorts consisting of the different types of estimates to categorize the cost growth, because the use of those mixes of cost estimates relate to different types of program structures, which might represent distinct populations with distinct cost growth patterns.

After looking at 318 programs across all of DoD, the cohort study results show that the PE and DE cohort has 30 percent RDT&E cost growth; the DE-only cohort has 25 percent RDT&E cost growth; the PE, DE, and PdE cohort has 35 percent procurement cost growth; the DE and PdE cohort has 25 percent procurement cost growth; and the DE-only cohort has 15 percent procurement cost growth. The sample sizes are 25, 140, 6, 53, and 94 respectively (Dameron, 2001:10). The NAVAIR group indicates that the "results are very tentative," but suggests that differences might exist in cost growth from one cohort to another. In particular, they point out that, in their study, "programs with a PDRR phase have more growth" (Dameron, 2001:11).

The NAVAIR study also looks at cost growth correlations between program phases and between the RDT&E and procurement appropriations. The study finds a significant correlation between RDT&E cost growth in the PDRR phase and RDT&E cost growth in the EMD phase and also finds "significant correlation between procurement growth during the EMD and production phases" (Dameron, 2001:14). Finally, it finds a significant correlation between appropriations such that, during EMD, when the RDT&E appropriation experiences cost growth, so does the procurement appropriation (Dameron, 2001:14).

As a third area of study, the NAVAIR group analyzes how program size affects cost growth. The team finds that the distributions of the high and low dollar programs are identical; however, "there is a trend of more high end extrema in the smaller size classes (though not statistically significant)" (Dameron, 2001:21). To explain the difference in the extrema, they reason that, "high risk programs may be terminated earlier if large, but tolerated if small" (Dameron, 2001:21). They find inferential statistics does not support a significant difference in the cost growth of programs based on the size parameters they study.

Next, NAVAIR studies the effects of the era in which an acquisition terminates and the cost growth occurs. As for the data, the team uses "DoD programs with DE only from the RAND 93 dataset, NAVAIR programs with DE only from the SAR 00 dataset, and NAVAIR programs with DE only from the Contract dataset (RDT&E only)" (Dameron, 2001:23). The team therefore has three separate data sets that they use, two of their own compilation and the RAND 93 dataset. The group studies the effects of two eras – pre-1986 and post-1986. They choose 1986 as a dividing point, because that year marks the last year of the Reagan arms buildup (Dameron, 2001:23). The team performs *t*-tests to determine if the two eras differ statistically. They find the following results:

- RAND 93: The means of programs through 1986 and those after 1986 did show a statistical difference for RDT&E, but not for procurement.

- SAR 00: The means of programs through 1986 and those after 1986 did show a statistical difference for procurement, but not for RDT&E.

- Contract: The means of programs through 1986 and those after 1986 did not show a statistical difference for RDT&E (Dameron, 2001:31).

The team concludes that their “analysis supports a decline in CGF over time” (Dameron, 2001:32). They mention that these results differ from previous studies perhaps because past studies have had too few data points in the newer era or because past studies have made bad choices for era division dates (Dameron, 2001:32).

The NAVAIR team further compares RDT&E cost growth in small programs (less than one billion dollars in RDT&E) as portrayed through the SAR 2000 data versus the NAVAIR contract database. This analysis concludes that the results from the two databases do not significantly differ (Dameron, 2001:34,38). They conclude that potential exists to use either database to study cost growth.

As a final area of research, the NAVAIR group studies differences between commodities and their relation to cost growth. The team looks at all three databases, but limits the data to 20 RAND 93 programs, 11 SAR 00 programs, and 21 contract data programs. They conclude that missile programs experience higher cost growth during RDT&E than either electronic or aircraft programs. Again, the scope of the NAVAIR study differs from the scope of our study, yet the study provides considerable insight into possible predictors for our research.

IDA Study

The Institute for Defense Analyses (IDA) performs an analysis on cost and schedule growth of tactical missiles and tactical aircraft in 1994 with the goal of finding patterns of cost growth and the reasons for the cost growth (Tyson, 1994:S-1). Within the group of 20 tactical missiles investigated, the IDA group finds that, "Programs took from 50 months to 137 months from Milestone II to initial operational capability" (Tyson, 1994:S-2). The study finds that only two of the 20 programs stay within their schedule, with one program slipping by as much as 180 percent, and that only two programs stay within budget, while the two worst performers exceed their budgets by a factor of two (Tyson, 1994:S-2). The researchers of IDA examine the characteristics of the programs with the highest and lowest schedule and cost growth (see results in Table 2 and Table 3) (Tyson, 1994:S-2). From their study, they find that:

> [Missile] programs that employed a high degree of concurrency, that had to be dual-sourced for technical reasons or that were dual-sourced at less than full rate, had high cost growth. In one case, the threat of competition appeared to reduce costs. (Tyson, 1994:S-2)

The results from aircraft programs do not vary as much. The authors of the study suggest closer management scrutiny and "protection from schedule stretch" as a reason for the more consistent cost growth in aircraft programs (Tyson, 1994:S-2). Two aircraft programs suffer from elongated production schedules, but do not experience high production cost growth as a result. The authors theorize that generally, stretching out the production program incites cost growth; however, in both of these aircraft cases the existence of other DoD contracts help cushion the impact of the adjusted schedules. The authors identify the F/A-18 as the program with the highest cost growth. They theorize that late engineering changes incite the high cost growth (Tyson, 1994:S-2).

Table 2. Characteristics of Programs with High and Low Schedule Growth in Development (Tyson, 1994:S-3)

Program	Percentage growth	Characteristics
Low Growth		
TOW 2	0%	Follow-on system
Sidewinder AIM-9M	1%	Follow-on system to fulfill goals of AIM-9L Learned from unrealistic estimate of prior system
MLRS	6%	Urgent program Competitive prototype Requirements/schedule tradeoff made in favor of schedule
High Growth		
Phoenix AIM-54A	94%	Problems resolved in development, not allowed to spill over into production Testing delays Delays in aircraft platform
Maverick AGM-65D/G	98%	Funding cut slowed development, allowed technology to catch up Prototype Vigorous testing program
AMRAAM	129%	Prototype showed infeasibility of approach High concurrency, urgent program Rushed testing
Sidewinder AIM-9L	148%	Urgent program, with fly-before-buy strategy Technical problems, with increased development quantity Joint service program, with technical disagreements
Sparrow AIM-7F	180%	Underestimation of technical difficulty (vacuum tube to solid state) Vigorous testing program

The study considers whether modification programs have lower cost growth than new start programs, and the results are as follows. The researchers find that the one aircraft in their sample that exists as a modification of a previous version of the aircraft does in fact experience low cost growth. The team finds that missile modification programs vary greatly in the amount of cost growth they experience. They cite the fact that most missile modifications affect the expensive guidance and control system of the missile as a possible reason for this inconsistency in missile modification program cost growth (Tyson, 1994:S-5).

Table 3. Characteristics of Programs with Low and High Cost Growth in Total Program (Tyson, 1994:S-4)

Program	Percentage growth	Characteristics
Low Growth		
MLRS	-10%	Competitive prototype Requirement lowered because of time urgency Mulityear procurement, low stretch
Maverick AGM-65A	1%	Total package procurement with low concurrency Vigorous testing program Low stretch
TOW 2	-4%	Urgent modification program Foreign Military Sales Low stretch
Sidewinder AIM-9M	10%	Learned from schedule problems in AIM-9L program Urgent program, took its lumps in development Low stretch
High Growth		
AMRAAM	84%	Prototype showed infeasibility of approach High concurrency, rushed testing Stretched program, dual-sourcing
Phoenix AIM-54C	89%	High concurrency Dual-sourced for technical reasons Five years qualifying for two years of competition Needed funding for next generation
Sparrow AIM-7M	100%	Competitive prototype, low cost growth in development Needed funding for next generation
Sidewinder AIM-9L	123%	Crash program Dual-sourced for technical reasons Production stretch

The researchers further find that the urgency of the program, the difficulty of the technology, the amount of concurrency, and the degree of testing all seem to affect cost growth in those programs studied (Tyson, 1994:S-5). From these results, the IDA researchers discover a relationship between cost growth and schedule growth in both the development and the production phases (Tyson, 1994:S-5). They find that quantity increases during development largely drive development schedule growth. The authors mention "the need to produce more items for testing than planned" as the reason for the increase in quantity (Tyson, 1994:S-6). The most obvious reason for producing more testing units is the need to repeat a failed test. Test failure, then, seems a reasonable candidate driver of schedule slip, and within the reasons for test failure (which IDA does not explore in depth, and nor shall we) might lie clues to program characteristics that would serve as good candidates for predictors of development schedule growth. The study also finds that whether a missile is an intercept missile and the length of the original schedule prove useful predictors of development schedule growth. From these relationships, the researchers go on to discover that, "Total program cost growth was related to total schedule growth, planned unit cost, and an intercept missile dummy variable" (Tyson, 1994:S-6). They calculate the equation for total estimated program cost growth as:

> TPCG= .7645 + (.3677*TSG) + (.1845*PUC) + (.2729*IMD)…where TPCG is total program cost growth, TSG is total schedule growth, PUC is planned unit cost in millions of 1994 dollars, and IMD is set equal to 1 for intercept missiles and 0 otherwise. (Tyson, 1994:S-6)

Using this equation, the researchers arrive at an adjusted R^2 of 0.500 and an SSE of 0.259. The coefficients have significance at a *p*-value of 0.04 (Tyson, 1994:S-6).

For aircraft programs, the researchers derive the following predictive formula for total program cost growth:

> TPCG = .3785 * ATS.2365 * EAV8B-.3962…where TPCG is total program cost growth, ATS is actual total schedule, and EAV8B takes the value 0 for the AV-8B and 1 for all other aircraft. (Tyson, 1994:S-7)

The researchers find this equation has an adjusted R^2 of 0.890 and an SSE of 0.053; the coefficients have significance below the 0.01 level (Tyson, 1994:S-6). The researchers conclude that, unlike the missile formula, which has an n of 20, the aircraft formula with an n of seven lacks enough data points to have usefulness as a predictive tool (Tyson, 1994:S-6). Both of these tools attempt to predict overall cost growth rather than a specific facet of cost growth.

In their analysis to discover the above-mentioned formulas for predicting total program cost growth, the IDA researchers consider several possible candidate independent variables. These candidate variables include schedule variables, program management variables, and program cost variables. Table 4, Table 5, and Table 6 list the predictor variables IDA considers. These candidate variables might prove useful as predictors for engineering cost growth.

Table 4. Candidate Independent Variables – Schedule Variables (Tyson, 1994:IV-2)

Variable	Notation	Definition	Development	Production and Total
Schedule Variables				
Planned development schedule	PDS	Planned time to develop the first version of the system, measured in months from Milestone II to IOC	X	
Actual development schedule	ADS	Actual time to develop the first version of the system, measured in months from Milestone II to IOC	X	
Development schedule growth	DSG[a]	Ratio of the actual development schedule to the planned development schedule	X	
Development schedule growth, predicted	DSGHAT	Predicted value of DSG in missile model (see section IV.B.)	X (missiles only)	
Planned production schedule	PPS	Planned time to produce the planned quantity of the system, measured in months from Milestone III to the end of production of the planned quantity		X
Actual production schedule	APS	Actual time to produce the planned quantity of the system, measured in months from Milestone III to the end of production of the planned quantity		X
Production schedule stretch	PSS	Ratio of the actual production schedule to the planned production schedule		X
Planned total schedule	PTS	Planned time to develop and produce the system, measured in months from Milestone II to the end of production of the planned quantity		X
Actual total schedule	ATS	Actual time to develop and produce the system, measured in months from Milestone II to the end of production of the planned quantity		X
Total schedule growth	TSG	Ratio of the actual total schedule to the planned total schedule		X

[a] DSG is also used as a dependent variable in the simultaneous model for missiles.

Table 5. Candidate Independent Variables – Program Variables (Tyson, 1994:IV-3)

Variable	Notation	Definition	Development	Production and Total
Program Variables				
Development quantity growth	DQG	Measure of growth in the development quantity	X	
Modification program	MOD	1 if the program is a modification program, 0 otherwise	X	X
Competition in full-scale development	CFSD	1 if competition (dual or multiple sources) was used in FSD, 0 otherwise	X (missiles only)	
Design-to-cost	DTC	1 if design-to-cost was applied, 0 otherwise	X	X
Total package procurement	TPP	1 if total package procurement was used, 0 otherwise	X (missiles only)	X (missiles only)
Incentives in full-scale development	IFSD	1 if contract incentives were used in full-scale development, 0 otherwise	X	
Prototype	PRO	1 if a prototype was developed before full-scale development, 0 otherwise	X	X
Competition in production	CPROD	1 if competition (dual or multiple sources) was used in production, 0 otherwise		X (missiles only)
Multiyear procurement	MYP	1 if a multiyear procurement contract was used, 0 otherwise		X
Fixed-price development	FPD	1 if fixed-price development was used, 0 otherwise	X	X
Full-scale development start	FSDST	The year of full-scale development start, used as a proxy for technological complexity	X	X
Concurrency	CONC	Percentage of test program remaining to be completed at Milestone III (see Reference [9])	X (missiles only)	X (missiles only)
Intercept missile dummy	IMD	1 if an intercept missile, 0 otherwise	X (missiles only)	X (missiles only)
IIR Maverick dummy	IIRMD	1 if an IIR Maverick (AGM-65D/G), 0 otherwise	X (missiles only)	X (missiles only)
AV-8B dummy	AV8BD	1 if an AV-8B, 0 otherwise	X (aircraft only)	X (aircraft only)
e AV-8B Dummy	EAV8B	*e* (= 2.71828) if an AV-8B, 1 otherwise	X (aircraft only)	X (aircraft only)

Table 6. Candidate Independent Variables – Total Cost Variables (Tyson, 1994:IV-4)

Variable	Notation	Definition	Development	Production and Total
Total Cost Variables				
Planned development cost	PDC	Planned cost to develop the system, measured in millions of FY 1994 dollars from Milestone II to the end of development of the first version	X	
Planned total cost	PTC	Planned cost of the total system at the Development Estimate, measured in millions of FY 1994 dollars from Milestone II to the end of production of planned quantity	X	X
Planned unit cost	PUC	Planned cost to produce a unit at the Development Estimate, measured in millions of FY 1994 dollars	X (missiles only)	X (missiles only)

Woodward Study

Peter Woodward specifies an elusive factor that can affect cost growth or the absence thereof in programs: the practice of hiding reserve funds within the budget

(Woodward, 1983:105). Indeed, if an estimate contains hidden reserve funds for uncertainties, then it has extra protection against cost growth than a program that does not have this hidden reserve. Finding such programs from SAR data proves next to impossible, making direct analysis of this phenomenon difficult. One must remember that program estimates may already include some reserve, whether hidden or overt, when developing a cost estimating methodology for risk from historical data.

RAND Study (1993)

In a 1993 study, RAND determines that inflation and quantity have the greatest effect on cost growth, but these two factors are part of the assumptions of a cost estimate initially, so RAND excludes them from cost growth for the purposes of their study. Although it might prove interesting to explore the historical distributions of these two factors with respect to cost growth, we adopt RAND's approach of excluding them from consideration.

The RAND study finds several other factors that relate to cost growth. Like the BMDO and NAVAIR studies, RAND considers program size. DoD categorizes acquisition programs according to how the programs compare to certain dollar thresholds. The higher dollar programs generally receive more management scrutiny than the lower dollar programs. More management scrutiny generally should translate into less cost increases due to mismanagement. Thus, one would expect to find a functional relationship between cost increases and the acquisition categories such that the programs in the higher acquisition categories have cost increases of a lesser magnitude. The authors of the RAND study offer another possible explanation for the difference in cost growth of the smaller programs, “R&D costs are a large portion of total costs and tend to incur more cost growth” (Drezner, 1993: 49).

The maturity of the program seems to also factor largely in the cost growth of a program. The RAND study notes that “on average, cost growth increases by 2.2 percent per year above inflation because of the effects of maturity.” RAND emphasizes the importance of these two factors above other factors in the statement, “Program size and maturity can dominate other factors affecting cost growth outcomes and so must be considered in both the analysis and the interpretation of results” (Drezner, 1993: 49). Therefore, these two factors represent prime candidates for predictor variables in a regression search of a cost risk factor function.

The RAND study elucidates the impact of new-start programs versus modification programs, finding that on average, the new-start programs experience more cost growth than modification programs. This stands to reason, and one should consider this distinction as a potential predictor variable. The RAND study also finds longer programs to have more cost growth than shorter ones. This simple linear relationship proves quite intuitive: each year brings the opportunity for more cost growth. “Of interest is that planned length and various measures of schedule slip are not related systematically to cost growth outcomes” (Drezner, 51: 1993).

Finally, RAND discovers that whether or not a program has a prototype effort has an opposite effect on cost growth than what the researchers at RAND expect:

> We compared the cost outcomes of prototyping and nonprototyping programs, expecting to find that a prototype development strategy contributes to cost control through reduction of uncertainty. Interestingly, programs that included prototyping had a relatively higher cost growth. This result may be due in part to the timing of the prototype phase within

> the context of the overall program schedule, since earlier prototyping makes data available earlier, thus potentially affecting the baseline cost estimate at the time of EMD start. Our results are consistent with this notion. It may also be true that prototyping was conducted for programs with relatively higher degrees of technical uncertainty, a hypothesis that deserves further exploration. (Drezner, 51: 1993)

From RAND's perspective, further research might help to determine if DoD uses prototyping on programs that are technically more risky than other programs. Basic DoD acquisition principles dictate that technically riskier programs use prototyping to reduce technical risk. The results of the RAND study do not necessarily defy reason: prototype programs, in fact, have more technical risk than non-prototype programs, and the prototyping probably does significantly reduce risk, but not necessarily to the extent so as to make a prototyped program have less cost growth than a non-prototyped program. In a cost growth model, we would then expect to use prototype or non-prototype as an explanatory variable for cost growth. Alternatively, we might use prototype or non-prototype as an indicator to determine some ordinal value of technical risk, which we might then use as a predictor variable in a cost model.

The RAND researchers conclude that "no single factor explains a large portion of the observed variance in cost growth outcomes" (Drezner, 52: 1993). This conclusion comes from a top-level, exploratory analysis of the total cost growth data. Whereas RAND finds no significant explanatory variables for overall cost variance, the possibility exists that breaking down cost growth into its components might uncover some significant explanatory variable. In addition, using multiple regression rather than simple linear regression might also prove useful in the search for significant explanatory variables.

Christensen and Templin Study

David Christensen and Carl Templin research cost growth using the Defense Acquisition Executive Summary (DAES) database and arrive at potentially useful findings in the search for predictors of cost growth [The DAES database contains contractor information organized according to the rules of Earned Value Management, a process by which the government monitors the cost and schedule performance of contracts against baseline figures] (Christensen, 2000:191). The researchers consider "hundreds of DoD defense acquisition contracts from 1975 through 1998" in a hypothesis testing scenario focused on the nature of management reserve (MR) budgets (Christensen, 2000:191). DoD characterizes the purpose of an MR budget as "a reserve for uncertainties related to in-scope but unforeseen work" (DoD, 1997:12). MR budgets, because they represent the contractors' assessment of risk for acquisition programs, can provide useful insight into the overall risk assessment that DoD uses in its budgeting process.

Christensen and Templin recognize that many factors affect the development of a contractor's MR budget, and that the "achievability of a budget depends on how the budgets are established" (Christensen, 2000:195). This gives the insight that overruns can vary depending on many factors, such as differing methods, differing abilities, and differing motivations of those who set the MR budgets (Christensen, 2000:193). A 1998 survey of 300 DoD risk analysis professionals supports this statement by displaying the variety of perspectives on risk analysis that exist within government and contractor circles (See Table 7) ("U.S. Aerospace Cost Risk Analysis Survey," 2000:23). In addition to the above, Christensen and Templin note that contractors should provide greater MR budgets for riskier projects. The authors go on to characterize the development phase of acquisition as more uncertain than the production phase, and they characterize price contracts as more uncertain than cost-reimbursement contracts (Christensen, 2000:196). From this awareness of the diversity of the risk analysis field, Christensen and Templin perform hypotheses testing to realize the following results:

> The amount of an MR budget is sensitive to contract category (cost-reimbursable versus fixed-price), and the managing service. With regard to contract category, the median MR percent on fixed-price contracts is significantly greater than the median MR percent on cost reimbursable contracts. This is consistent with the expectation that contracts with more risk to the contractor have a larger MR budget. We do not know why MR budgets differ across the three services. Possible explanatory factors include differences in the weapon systems purchased by each service, and the contractors that build the systems. (Christensen, 2000:204)

With regard to the acquisition phase, the researchers do not find that the MR budget differs between production and RDT&E contracts (Christensen, 2000:202).

Table 7. Unexpected Findings ("U.S. Aerospace Cost Risk Analysis Survey," 2000:24)

- 27% of analyses perform the risk assessment separately from the cost estimate.
- 26% of program managers do not accept risk assessment at all, not even "slightly."
- 32% of the risk assessments do not involve Finance or Estimating.
- 38% of cost risk analysts have received no training, either formal or informal.
- 44% of risk ranges are intuitive judgments, without historical data or guided-survey.
- 69% of variable distributions are triangular.
- 18% of unfavorable assessments are ignored, as managers "stay the course."

Wilson Study

A 1992 study of the DAES database by Brian Wilson provides more insight into possible predictor variables for cost growth. (Wilson, 1992:42). Wilson uses hypothesis testing on a database of 109 contracts spanning the period from 1977 to 1991. Wilson discovers two trends that provide insight into the current study. First, he finds at the 85 percent confidence level that "cost overruns tend to worsen as a contract progresses toward completion" (Wilson, 1992:81). Secondly, Wilson finds that the pattern of cost overruns over time depend upon certain program characteristics (Wilson, 1992:81).

Wilson finds the following characteristics to explain significant differences in cost growth at the 85 percent confidence level: service, contract type, system type, and program phase. For type of service, Wilson finds significant differences between Army and other service programs [Wilson considers Marine Corps programs as Navy programs]. For contract type, Wilson finds significant differences between cost plus contracts and fixed price contracts. Wilson finds significant differences between the air based, sea based, and land based systems. His initially chooses those system types for testing in order to "minimize the number of system types used," and order to distinguish between the level of "required reliability for each" (Wilson, 1992:48). He uses as an example the fact that the consequences of failure of a land-based jeep place minimal risk upon a user compared to the consequences of failure of an aircraft. Implicitly he seems to assume that the level of reliability required relates to potential for cost overrun. Finally, Wilson finds significant differences between overruns in development contracts and those in production contracts. These results provide clear possibilities with which to explore possible predictors of cost growth within the SAR database.

Terry and Vanderburg Study

In another analysis using 321 defense contracts from the DAES database, Mark Terry and Mary Vanderburg analyze contractor estimates at completion (EAC) and their relationship to the contractor actual Cost at Completion (CAC) using hypothesis testing. The researchers null hypothesis is that "Cost at Completion is bounded below by the Cost Performance Index (CPI)-based EAC and above by the Schedule Performance Index (SCI)-based EAC" (Terry, 1993:23). They resolve at the end of their analysis that they should reject the null hypothesis (Terry, 1993:59).

Of interest to our study, the researchers test for sensitivity of their results to the following attributes: "Index Type (cumulative, six-month and three-month), Contract Completion Stage, Program Phase, Contract Type, Branch of Service, System Type, Major Contract Baseline Changes, and Management Reserve" (Terry, 1993:59-60). The sensitivity analysis performed on the Terry cost growth study, shows that the amount of cost growth in the 321 defense contracts depended to some degree on the attributes mentioned above, with the exception of management reserve (Terry, 1993:60). Thus in our study, these same attributes might prove useful as independent variables.

Obringer Study

In a study by Thomas Obringer in 1988, the author studies overall cost growth in the defense aerospace industry during the period 1980 to 1986 (Obringer, 1988:5). He gathers data from Business Management Information Reports (BMIR) for 16 contractor plants, which at the time comprise 32 percent of the industry's sales. In his study, he uses hypothesis testing to discover that from 1980 to 1986 aerospace industry costs do not rise in real terms. He also discovers a similar trend in overhead rates. The first finding conflicts with similar studies performed in the 1960's and 1970's on the aerospace industry, where increasing costs emerged as the dominant trend. Obringer notes that increased defense spending and decreased excess capacity characterize the period of his study, alluding to a possible reason for the difference in the results from previous eras (Obringer, 1988:84-86). The Obringer study along with the studies of the previous two decades suggests that era might affect the results of our study. Though we will not seek to explain the effects of era on cost growth, we limit the effects of era on our results by limiting the timeframe of our study to a single decade (1990-2000).

An additional observation from the Obringer study, the stability of the composition of aerospace firm costs, also provides useful insight into cost growth. Obringer notices that his study reveals a cost composition remarkably similar to the studies of the 1960's and 1970's (Obringer, 1988:84). Generally speaking, all three studies show that components of contractor costs have raw materials as the highest percentage of total costs and overhead costs as the next highest, the two respectively comprising about half and a third of total contractor costs (Obringer, 1988:79). Since these two components of company cost consistently comprise over 75% of total costs, the root causes for most cost growth likely lie within them. Of course, one must temper the potential applicability of Obringer's findings by the narrow, aerospace-industry focus of his study.

<u>Singleton Study</u>

Pamela Singleton, in her thesis, investigates the causes of cost growth in large and small acquisition programs initiated by the then Aeronautical Systems Division from 1980 through 1988 (Singleton, 1991:7). Singleton measures cost growth as the difference between the most probable cost (MPC) estimate (calculated for use during source selections) and the most current estimate of the program (for completed programs, this is the actual program cost). Singleton performs a literature review and solicits a group of five cost analysts to come up with a reasonable list of factors affecting cost growth in weapon systems. The panelists rank the factors in order of effect on cost growth, culminating in the following top three list: "technical risk, configuration stability, and schedule risk" (Singleton, 1991:75). Singleton then collects data on 16 programs from the Aeronautical Systems Division to test whether or not cost growth correlates to the three factors above in this subset of programs. She finds that:

> When the effect of all three factors are considered together in the development effort, configuration stability tends to have more influence on cost growth than the other factors. The analysis suggests that significant cost growth should be expected if the program is operating in an environment with low configuration stability and high schedule risk. Though high configuration stability does not guarantee minimal cost growth, the cost growth experienced in these programs tends to be less on average than those with low configuration stability. (Singleton, 1991:76)

Singleton also finds that "reducing technical risk will not significantly decrease cost growth if there is a high probability that the schedule will slip six months or more" (Singleton, 1991:76).

In the production stage, Singleton finds similar results. When she removes the other two factors from the scenario, Singleton finds that configuration stability greatly influences cost growth; however, when one considers all factors together, the stability of the system's configuration plays no significant role in driving cost growth (Singleton, 1991:76-77). Schedule risk, on the other hand, carries great weight in driving cost growth when including all three factors. She states that, "in all instances where the schedule risk was high, the cost growth exceeded eighteen percent" (Singleton, 1991:76).

Although, the limited sample size hinders the ability to generalize Singleton's results, the care that she took in determining which variables to test provides for optimism that her results will benefit our research. The SAR database does not systematically include information regarding the three risk factors of Singleton's study, but some proxy for these factors may provide reasonable predictability in a cost growth model.

As a side note to the Singleton study, DoD describes the following as "relevant sources of risk" for cost analysts to consider:

> …design concept, technology development, test requirements, schedule, acquisition strategy, funding availability, contract stability, or any other aspect that might cause a significant deviation from the planned program. Any related external technology programs (planned or on-going) should be identified, their potential contribution to the program described, and their funding prospects and potential for success assessed. This section should identify these risks for each acquisition phase (DEM/VAL, EMD, production and deployment, and O&S). (Department of Defense, 1992:9)

The cost estimator must describe these sources of risk in the CARD for each program submitted to the CAIG for review (Department of Defense, 1992:3). This description of potential risk inciters gives clues as to potential drivers for program cost growth.

Eskew Study

In an effort to find the true rate of cost growth of fighter aircraft over time, Henry Eskew runs a linear regression of 17 tactical aircraft from 1950 through 1980 (Eskew, 2000:210). He normalizes his data for production quantity by using the estimated 100th production unit cost, and he normalizes his data for inflation by applying the appropriate

DoD inflation indices to convert his data to constant year (CY) 1990. Using the logarithm of cost as his response variable, he finds weight, speed, production rate, and time as statistically significant predictor variables that explain "more than 90 percent of the variation in cost" (Eskew, 2000:211-212). He also determines that, as a sole predictor, time explains about 40 percent of the cost variation, hinting at its possible significance as a predictor for our study (Eskew, 2000:211). In fact, all four of the predictors he finds might prove useful areas of exploration in search of predictors for our study.

Although useful, one must note the limitations of the Eskew study's applicability to this study. First, Dr. Eskew looks at a limited amount of data from a limited perspective. He only considers tactical aircraft in his search for predictors, and he only has 17 data points. In addition, his research spans the period from 1950 through 1980, whereas the current research spans the period from 1990 to 2000. This limits the confidence that one can have in the applicability of his results to the research at hand. Secondly, the perspective of cost growth in his study differs from the perspective of cost growth for our research. Dr. Eskew's research seeks to explain cost growth as overall increases in unit cost measured from previous programs over time. (Eskew, 2000:209). The research of this thesis seeks to explain the growth of cost from the initial Development Estimate as recorded in the SAR database. Thus, the results of his analysis give insight into possible predictors of cost growth for the purposes of this study, but only to the extent that the predictors he finds for cost growth from program to program relate to cost growth from the Development Estimate to the actual program costs.

In the same research paper, Dr. Eskew seeks to dispel the myth that "no systematic relationship exists between the characteristics of an aircraft program and the length of its development cycle" (Eskew, 2000:210). He uses the same normalization techniques mentioned earlier for inflation and quantity; however, he includes different aircraft, adding non-tactical fixed wing aircraft, and removing non-fixed wing aircraft (Eskew, 2000:214). The results of his 18 data-point regression show that unit flyaway cost predicts approximately 60 percent of the variance in the length of the development program: this predictive ability increases to 70 percent when a dummy variable is added indicating whether or not a program has inherited a significant amount of technology from a previous program (Eskew, 2000:214-215).

Chapter Summary

In this chapter, we document many studies that query different databases using various statistical methods in the quest to explain cost growth in DoD acquisition. From these studies, we derive the following general list of predictor variables that we will pursue in our research: program size, physical type of program, management characteristics (military and contractor), schedule characteristics (maturity and concurrency measures), and other characteristics mentioned in the literature review. None of the historical studies deals directly with cost growth in the RDT&E budget from engineering changes in EMD. Yet, from the results of these studies, we might gain the insight required to successfully find predictors of engineering cost growth in the current study.

III. Methodology

Chapter Overview

This chapter describes the process by which we conduct this research. We begin by reviewing our literature results, which provide clues that help us select possible predictors of cost growth. The literature results also form a backdrop of knowledge in which we critique the results of this research. We next assess the data source and describe the process by which we collect and compile the data. Finally, we describe the exploratory data analysis and regression techniques that we use.

Literature Synopsis

The 1993 RAND study based on SAR data serves as the cornerstone of the literature review. The RAND corporation accomplishes a descriptive statistical analysis on overall cost growth (normalized for inflation and quantity changes) in both procurement and RDT&E dollars, from which we form general impressions about cost growth as it relates to different programmatic characteristics. We go on to analyze several studies pertaining to the broad areas of cost growth and risk analysis. We fail to find a study that shares the narrow focus of our study - RDT&E cost growth in the EMD phase due to engineering changes. Thus, our study differs from all other studies in one very important way: our study allows for the possibility that a group of candidate predictors may vary in either their ability to predict or the degree to which they predict the seven different SAR categories of cost growth. This difference limits the applicability of our literature review somewhat, but by no means renders null its emolument: the review still provides useful clues toward our purpose. We limit ourselves to predictors that we can find within the SAR data. Consequently, some of the clues we find in our literature review remain fertile ground left for future researchers to explore.

Search for Predictors of Cost Growth

From the past research, we identify possible predictor variables for the current study in cost growth. Ideally, we detect logical causality between predictor and response variables; however, apparent causal relationships need not exist for inclusion as a candidate variable. We need only suspect a reasonable prediction possibility for consideration as a candidate predictor.

As an example of predictability without apparent causality, consider the independent variable *whether a program had a PDRR phase.* This independent variable would seem to have no causal relationship with *whether a program in the EMD phase will have cost growth*. Contrarily, we suspect the *level of program uncertainty at the start of EMD* does have a causal relationship with *whether a program in the EMD phase will have cost growth*. However, we have no way of determining from our data the *level of program uncertainty at the start of EMD*. We recognize that *whether a program had a PDRR phase* logically correlates to the uncertainty level. Thus, we use *whether a program had a PDRR phase* as a proxy for the *level of program uncertainty at the start of EMD*.

In our search for predictors, we keep in mind that the estimator must either know or be able to estimate the predictors chosen at the time the program office accomplishes the DE. In other words, a candidate variable might accurately predict cost growth, but if the cost estimator has no idea of the value of that candidate variable at the time he produces the estimate, he cannot use it to produce a cost estimate of the response variable. Thus, a model that we produce must not include such recondite variables.

Finally, an estimator must easily understand the relationship between the predictor variables and the response variables in any models we discover. If the estimator does not understand the variables, two problems arise. First, the estimator might lack faith in the model, causing him to discredit its results. Second, even if the estimator supports the model, he will not have the ability to support it in the event if falls under management scrutiny. Thus, the predictors we find do not have to demonstrate an apparent causal relationship with the response variables, but they must have some logical tie to the response variables that the estimator can easily understand, and they must be available at the time of the estimate.

Database

We use cost variances and other information as recorded in the Selected Acquisition Report (SAR) database for this analysis. The SAR data records cost variances in base year as well as then year dollars. We use the base year dollars for analysis, since these dollars exclude estimated inflationary effects. This format facilitates conversion of the various base years of individual estimates into a single base year, making possible easy comparison across programs. The SAR records cost variances in seven different categories:

- Economic: changes in price levels due to the state of the national economy
- Quantity: changes in the number of units procured

- Estimating: changes due to refinement of estimates
- Engineering: changes due to physical alteration
- Schedule: changes due to program slip/acceleration
- Support: changes associated with support equipment
- Other: changes due to unforeseen events (Drezner, 1993:7)

In addition to these categories, the SAR also provides the total cost variance - the sum of the above seven variances. The RAND study of 1993 analyzes the total cost variance, whereas this thesis focuses on cost variance due to engineering changes specifically. The RAND researchers only study positive cost variances (i.e., cost growth). Similarly, the study at hand focuses on cost growth, but does consider zero and negative cost variances to a degree, thus we collect all cost variance data, whether zero, positive, or negative.

The SAR database contains a variety of programmatic information from major defense acquisition programs from all military services. This information includes historical, schedule, cost, budget, and performance information for the life cycle of the program. Only programs that meet the dollar thresholds or that have the Congressional interest making them ACAT IC or D programs have files in the SAR database (Knoche, 2001:1). The ACAT criteria change over time, but the programs listed in the SAR files consistently represent programs with high-level government interest. In some cases, the information we desire from the SAR database has a security classification. For security reasons, we do not use this information in the compilation of our database. Thus, the subset of the SAR database we use represents a compilation of the programmatic details of some (but not all) of the most important DoD programs.

Our research is not the first using the SAR; in the early 1990's, RAND researches the SAR and produces a modified SAR database. This database contains selected information from individual SARs in spreadsheet format. Unfortunately, the RAND spreadsheets do not break cost growth into its seven parts as mentioned above. In addition, the latest entries in the RAND database date back to the early 1990s. Finally, the RAND database lacks adequate information on many of the predictor values that we wish to investigate. All of these shortcomings make the RAND database useful only as a verification tool for part of the data collection effort.

The SAR Database as a Source of Historical Data

According to RAND, researchers of cost growth commonly use the SAR database to conduct their research (Hough, 1992:v). RAND notes that while the government has continually improved the quality and consistency of information included in the SAR database, the database still has numerous "pitfalls" that a cost analyst must attenuate in order to maximize the validity of analyses based on the SAR (Hough, 1992:v). Those problems most prevalent follow:

- Failure of some programs to use a consistent baseline cost estimate
- Exclusion of some significant elements of cost
- Exclusion of certain classes of major programs (e.g., special access programs)
- Constantly changing preparation guidelines
- Inconsistent interpretation of preparation guidelines across programs
- Unknown and variable funding levels for program risk
- Cost sharing in joint programs

- Reporting of effects of cost changes rather than their root causes (Hough, 1992:v)

The SAR provides some consistency in the reporting of programmatic data, but as RAND describes, "…although the basic content of the SAR sections is established by DoD Instruction 7000.3, interprogram comparisons can be complicated by the fact that specific details vary" (Hough, 1992:4). In addition to differences in specific details, the guidelines themselves change over time, providing a further source of inconsistency (Hough, 1992:4). Despite possible difficulties with the data, RAND recognizes the SAR as "the logical source of data for calculating cost growth on major procurements" (Hough, 1992:9).

For cost estimating purposes, the cost analyst should normalize the cost growth for inflation and quantity changes, because these can have a large effect on the cost growth, and reliable methods exist with which to make these adjustments (Hough, 1992:10). The SAR format devotes two of the seven cost variance categories to capture these adjustments. Thus, we have no such adjustments to make on the data we collect from the SAR engineering cost growth category.

According to RAND, the cost analyst must decide from which baseline to measure cost growth. The SAR offers three different baselines, the planning estimate (PE), the development estimate (DE), and the production estimate (PdE). These estimates occur before the start of Milestone I, II, and III, respectively. The RAND study mentions that cost estimates performed later in the product's life cycle more accurately reflect the program cost. This observation by RAND holds true, because program uncertainty drives the accuracy of cost estimates, and as programs progress uncertainties become certainties. It follows that cost growth increases as measured from the PE versus the DE, and it increases as measured from the DE versus the PdE (Hough, 1992:10-11). Figure 3 presents the relationships between the different baseline estimates and the acquisition phases, adding the additional categorization of funding appropriation. Thus, one can consider several different measures of cost growth: growth of procurement dollars from the PE; growth of procurement dollars from the DE; growth of procurement dollars from the PdE; growth of RDT&E dollars from the PE; and, the cost growth our study researches, growth of RDT&E dollars from the DE.

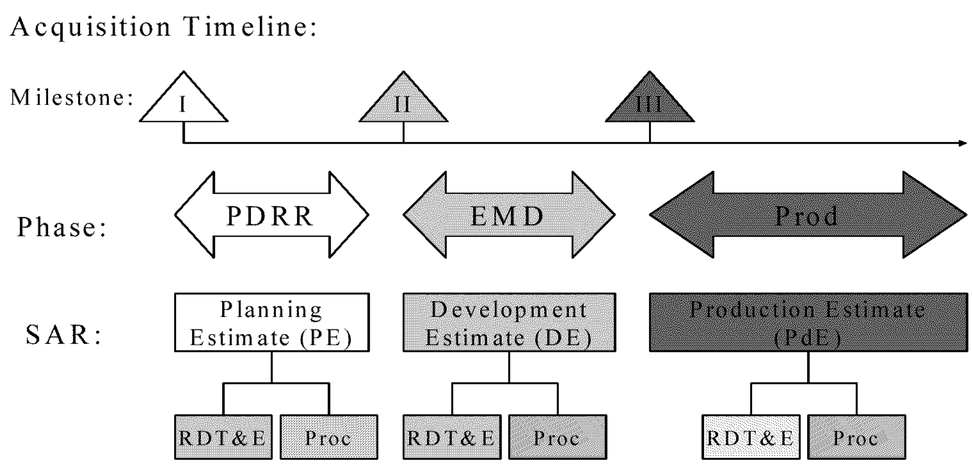


Figure 3. Introduction to SARs-SAR Types (Dameron, 2001:4)

RAND defines cost growth as "the difference between the most recent or final estimate of the total acquisition cost for a program and the initial estimate" (Hough, 1992:10). This definition applies to program cost growth as measured from the first estimate made (PE, DE, or PdE depending on the program structure) through the end of the program. Since our research investigates cost growth only in the EMD phase, we consider as our initial estimate the DE. This echoes the method used in the NAVAIR study mentioned in Chapter II of this thesis, whereby they calculate different cost growth factors for each acquisition phase. The PE, DE, and PdE respectively serve as denominators of those cost growth factors. Our study ultimately seeks to predict cost growth in the form of a factor to apply to a cost estimate. As such, we combine the RAND and NAVAIR philosophies. For our calculations, we compute percent engineering cost growth by first calculating the difference of the current estimate minus the DE. Then we divide the result by the DE. The SAR data contains all the necessary information to make these calculations.

The Baseline Problem

RAND notes that even though an analyst selects a certain baseline from which to measure cost growth, that baseline might not represent a consistent measure across different programs for two reasons. First, rebaselining might occur (the program office accomplishes a new baseline in the middle of an acquisition phase). This new program estimate retains the name PE, DE, or PdE (as appropriate), making it indistinguishable from a program that does not have an estimate rebaseline. If an analyst does not choose the correct DE from which to measure the program cost growth, the rebaselining will understate cost growth. RAND mentions that this happens infrequently. Evolutionary model changes provide a second reason for inconsistency in the baseline. Evolutionary model changes occur when a program's "configuration has been modified so much that current models only remotely resemble what was originally estimated." These changes prove difficult to normalize out of the SAR data (Hough, 1992:12-14).

Exclusion of Certain Program Costs

The RAND researchers identify the fact that the SAR excludes certain program costs. The SAR does not require all categories of costs relevant to an acquisition program. Operating and Support Costs at the time of this RAND study has no place in the SAR. This practice has since changed, but enough time has not elapsed so that such a change permeates the entirety of any extensive cost database built from the SAR data. A second exclusion from the SAR, technical deficiency, prevents the precise measurement of a deviation from the baseline cost estimate. In order to have a precise measure of this deviation, any technical tradeoffs made would need quantification and inclusion in the SAR. Third, contractor-borne expenses do not appear in the SAR. These expenses occur when a contractor invests his own money in a project or when certain fixed-price contracts with contractors prevent the disclosure of cost increases within contract limits. As a fourth example of excluded costs, RAND mentions that in some cases, spare parts, simulators, and other types of costs that have a clear link to the program do not receive recognition in the SAR (Hough, 1992:12-47).

Additionally, RAND speaks of the practice of postponing the reporting of cost growth as "closely related to the problem of unrecognized costs." Program managers will postpone cost reporting until after a significant milestone decision, which results in cost growth reporting in the incorrect program phase. As a final note, the RAND researchers relate that the ties a program estimate must maintain to the annual budget can cause a delay in the reporting of certain cost growth information if the president's budget does not receive Congressional approval in time for inclusion in the annual, December SAR. For most, if not all, of these excluded costs, RAND advises that reasonable means do not exist by which to normalize the SAR data; the estimator should however, be aware of the possible effects of their exclusion on the accuracy of the data (Hough, 1992:12-17).

Incomplete and Evolving Database

Commonly, an estimator performs analysis based on a small portion of the SAR database. RAND notes that the estimator must take care to ensure that the sample pulled from the SAR represents the population that the analysis seeks to characterize. RAND states, "...quality studies on cost growth should identify what portion of the total SAR population is included and why the sample is representative of the whole or is satisfactory for meeting the study objectives." Additionally, the SAR database excludes a large portion of defense programs – those programs of high security classification. This exclusion makes the entire SAR database incomplete to start with, and worse, these excluded, high-security programs represent the bulk of the programs pushing the envelope of modern technology (Hough, 1992:16-18).

Inconsistency in SAR Preparation Guidelines and Techniques

In order to improve the quality of the SARs, Congress continuously changes the preparation guidelines. Although these changes often have no impact on cost data, occasionally they do have a significant impact. These impacting changes may improve the quality or content of the data, but negatively affect the uniformity of the database such that comparisons over time must receive scrutiny. Further aggravating this point, not all organizations adopt the changes at the same time; RAND observes that, "after a major change, consistency among SARs is not ensured until all programs with current reporting begin under the same set of rules" (Hough, 1992:19-20).

RAND further notes that pressures to bias the SAR data exist on certain levels such that preparers might attempt to skew the data in some manner. This detracts from the accuracy of the data, possibly confounding any sort of analysis attempted on the database. As a program matures, the amount of unknowns in a program decreases, which in turn decreases the places in a program budget where room exists for stretching the reasonableness of assumptions in such a way as to bias the estimate (Hough, 1992:20-21).

Unknown and Variable Funding Levels for Program Risk

Cost estimators include monetary padding for risk within their estimates. Because of the instability of the acquisition-funding environment, Congress and the services often take money from one program to fund another. To avoid becoming a victim of this budgetary cannibalization, programs will often covertly include their management reserve funding within another budget line item. Thus, the SAR includes estimates for risk, the methodologies for which vary from service to service and from estimator to estimator. Quantification of these risk estimates proves impossible, making normalization of the database, in this regard, impossible as well (Hough, 1992:21).

Cost Sharing in Joint Programs

In joint programs, estimators can apply costs for investment to one program or spread the costs across all participants in some way. No guidelines exist for a single method in handling such situations. The lack of guidance causes inconsistency in the reporting of large portions of cost in joint programs, creating innate inaccuracies in cost growth analysis across joint programs (Hough, 1992:22).

Reporting Effects of Cost Changes Rather Than Root Causes

Although the SAR has a section for change explanations when cost growth occurs, the SARs do not systematically disclose the "root causes" of cost growth. An analyst might forage through other sections of the SAR to look for clues that point to the "root causes," but the results of such searches may prove questionable. This weakness hinders the ability of the SAR to provide analysts with drivers of cost growth (Hough, 1992:23).

The RAND Corporation thoroughly assesses the limitations within the SAR database. These limitations do not prevent us from discovering a cost model that provides benefit to the end user. In fact, as databases go, a database built from SARs has advantages to it that other databases do not. First, it conforms to a strict reporting format, providing consistency to the data. Second, those who create SAR reports receive annual SAR training, which adds to the consistency of the data (Knoche, 2001:2.B.3.2). Third, because SARs go before Congress, the level of scrutiny that SARs receive in the review process bolsters both the consistency and accuracy of the documents. Databases in general contain inaccuracies, but a database built from SAR data arguably withstands scrutiny better than most.

Data Collection

Security classification poses an obstacle to data collection. The SAR database contains some sensitive information. Indeed, some information that might prove useful to this research effort is sensitive on certain programs. We exclude data that has a security classification; thus, the research database has incomplete information for some programs and excludes other programs altogether (when the entire file is classified).

The SAR database contains thousands of individual files representing a variety of programs from each service over a variety of years, and each SAR report contains a great deal of information that can potentially prove useful for this research effort. In this situation, we narrow the scope of the data we collect. Because of the broad nature of some of the goals in the research effort, such as seeking the effects of joint program management in cost variance, the data collection effort does not exclude the collection of files based on branch of military service or program type. Instead, the data collection starts with the most recent SAR data available and works backwards, collecting an initial, broad collection of data points to arrive at preliminary research results.

Additionally, we desire the most current information to capture recent trends. Thus, we start our collection effort with the latest SARs available and work backwards in time until we have a sufficient number of data points to support a statistically significant regression. Specifically, the latest SARs at our disposal date within the summer of 2000. Thus, we start with those SARs and work backwards through the entire 1990 collection. As discussed earlier, we exclude those SAR files that have preventive security classifications. We also only include one SAR for each program – the latest. This ensures we have independence of data points. Further, since this research effort only concerns cost growth in the EMD phase, we include only SARs for which the DE serves as the baseline estimate. Once we determine which files to collect, we decide what information within the file might prove useful for predicting engineering cost increases.

Not only do we have to determine which information from the SAR files to extract, but also in what form the data will prove useful. In some cases, we perform mathematical operations between data in the SAR files to arrive at a possible predictor. The predictors RAND uses in their 1993 study provide guidance as to the form of some of the predictors we use. In other cases, we find similarities between SAR files and categorize those files accordingly. In a few cases, we seek outside sources to fill in gaps of information that the SAR leaves out. In these ways, we not only narrow down the information within the SARs, but also create additional information leveraging from data within the SARs.

Exploratory Data Analysis

Before data analysis, we expect to find a continuous distribution of data upon which we can perform multiple regression analysis in order to find a sufficient predictive formula. However, after collecting and analyzing the data, we find the response variable to have a mixed distribution. About half of the distribution is continuous, while the other half is massed on one value, zero. This mixed distribution scenario generally calls for splitting the data into two sets.

This splitting of the data logically follows from the incongruity between the two distributions. In a continuous distribution, the probability of obtaining a specific value is approximately zero. Such a probability does not accurately reflect the fact that many of the points in our data fell directly on zero. For the discrete distribution, we use logistic regression, and we use multiple regression analysis for the continuous distribution. Thus, we develop a logistic regression model to predict whether or not a program will have cost growth from a full data set, and we develop a multiple regression model from only those programs that had cost growth to predict the amount of cost growth we expect. For comparison purposes, we decide to pursue a single-step multiple regression model as well. This serves to ascertain what would occur if one overlooked the mixed distribution and attempted an estimation of the mean cost growth.

In addition to the mixed distribution, we find that a few of the programs have negative engineering cost variance. A user would not realistically assign negative values to cost growth in an estimate; however, we consider the negative values in the creation of the single-step multiple regression model. For the logistic regression portion of our analysis, we convert all negative cost growth to zero cost growth.

Before we start the regression analysis, we set apart approximately 20 percent of our data for validation purposes and sensitivity analysis. We use the random number generator in Microsoft® EXCEL (Microsoft, 2000) that uses a uniform distribution to choose which data we set aside. Before performing regression, we must also choose the response and candidate predictor variables.

Response Variables

As mentioned in Chapter I, this research seeks to find predictors of cost increase due to engineering changes. The SAR includes two main categories of engineering cost growth - increase in the research and development budget and increase in the procurement budget. Additionally, adding the two gives the total engineering change increase. Consequently, three possible response variables arise from the SAR data. We find it necessary to consider RDT&E and procurement separately, because certain predictor variables might work contrary to RDT&E cost growth versus procurement cost growth; therefore, we discard as a possible response variable the total cost increase due to engineering changes. Further, in the interest of time, we choose to study only the RDT&E increases.

We concern ourselves with two different response variables, one that indicates if cost growth will occur and another that expresses the degree to which cost growth occurs. The first of the two, we express as a binary variable where the value '1' means that we estimate a program will have engineering cost growth in RDT&E dollars, while the value '0' means that it will not. We call this variable *R&D Cost Growth?*.

In order to make the model as useful as possible, we decide that the second response variable should have the form of a percentage, rather than a dollar amount. The percent format applies well to programs with both large and small acquisition costs, whereas the dollar amount format might require us to force a program size variable into the model for the results to intuitively make sense. For example, a model with length of EMD and maturity from milestone III decision might produce a predicted engineering cost growth of 50 million dollars for both a 100 million dollar program and a 5 billion dollar program. Although this might be a valid result statistically, it might prove difficult for program managers to put into context. Thus, we strive to find a model to predict percent change in RDT&E cost due to engineering changes. As discussed earlier in the chapter, we use the DE as the denominator of the percentage. We call this second response variable *Engineering %*.

Predictor Variables

Several possible predictor variables exist within the SAR data. We aim to create a tool for cost estimators to create more realistic estimates, so the inputs for such a tool must be available to the estimator at the time of the estimate. However, we do not exclude variables from our analysis that do not meet this availability criterion. Rather, we analyze those variables to discover if predictive ability exists in the hopes of finding some correlated variable that the estimator might have available at the time of estimate creation.

The predictor variables we extracted from the SAR fall into five broad categories: program size, physical type of program, management characteristics, schedule characteristics, and other characteristics. Within these broad categories, we create two levels of subcategories. We list the predictor variables below by category and subcategories along with a short description of the subcategories that includes explanation of ambiguous elements where necessary:

Program Size Variables

- *Total Cost CY $M 2001* – continuous variable which indicates the total cost of the program in CY $M 2001
- *Total Quantity* – continuous variable which indicates the total quantity of the program at the time of the SAR date; if no quantity is specified, we assume a quantity of one (or another appropriate number) unless the program was terminated
- *Prog Acq Unit Cost* – continuous variable that equals the quotient of the total cost and total quantity variables above
- *Qty during PE* – continuous variable that indicates the quantity that was estimated in the Planning Estimate
- *Qty planned for R&D$* – continuous variable which indicates the quantity in the baseline estimate

Physical Type of Program

- Domain of Operation Variables
  - *Air* – binary variable: 1 for yes and 0 for no; includes programs that primarily operate in the air; includes air-launched tactical missiles and strategic ground-launched or ship-launched missiles
  - *Land* – binary variable: 1 for yes and 0 for no; includes tactical ground-launched missiles; does not include strategic ground-launched missiles
  - *Space* – binary variable: 1 for yes and 0 for no; includes satellite programs and launch vehicle programs
  - *Sea* – binary variable: 1 for yes and 0 for no; includes ships and ship-borne systems other than aircraft and strategic missiles
- Function Variables
  - *Electronic* – binary variable: 1 for yes and 0 for no; includes all computer programs, communication programs, electronic warfare programs that do not fit into the other categories
  - *Helo* – binary variable: 1 for yes and 0 for no; helicopters; includes V-22 Osprey
  - *Missile* – binary variable: 1 for yes and 0 for no; includes all missiles
  - *Aircraft* – binary variable: 1 for yes and 0 for no; does not include helicopters
  - *Munition* – binary variable: 1 for yes and 0 for no
  - *Land Vehicle* – binary variable: 1 for yes and 0 for no
  - *Ship* – binary variable: 1 for yes and 0 for no; includes all watercraft
  - *Other* – binary variable: 1 for yes and 0 for no; any program that does not fit into one of the other function variables

Management Characteristics

- Military Service Management
  - *Svs > 1* – binary variable: 1 for yes and 0 for no; number of services involved at the date of the SAR
  - *Svs > 2* – binary variable: 1 for yes and 0 for no; number of services involved at the date of the SAR
  - *Svs > 3* – binary variable: 1 for yes and 0 for no; number of services involved at the date of the SAR
  - *Service = Navy Only* – binary variable: 1 for yes and 0 for no
  - *Service = Joint* – binary variable: 1 for yes and 0 for no
  - *Service = Army Only* – binary variable: 1 for yes and 0 for no
  - *Service = AF Only* – binary variable: 1 for yes and 0 for no
  - *Lead Svc = Army* – binary variable: 1 for yes and 0 for no
  - *Lead Svc = Navy* – binary variable: 1 for yes and 0 for no
  - *Lead Svc = DoD* – binary variable: 1 for yes and 0 for no
  - *Lead Svc = AF* – binary variable: 1 for yes and 0 for no
  - *AF Involvement* – binary variable: 1 for yes and 0 for no

- *N Involvement* – binary variable: 1 for yes and 0 for no
- *MC Involvement* – binary variable: 1 for yes and 0 for no
- *AR Involvement* – binary variable: 1 for yes and 0 for no

- Contractor Characteristics
  - *Lockheed-Martin* – binary variable: 1 for yes and 0 for no
  - *Northrup Grumman* – binary variable: 1 for yes and 0 for no
  - *Boeing* – binary variable: 1 for yes and 0 for no
  - *Raytheon* – binary variable: 1 for yes and 0 for no
  - *Litton* – binary variable: 1 for yes and 0 for no
  - *General Dynamics* – binary variable: 1 for yes and 0 for no
  - *No Major Defense KTR* – binary variable: 1 for yes and 0 for no; a program that does not use one of the contractors mentioned immediately above = 1
  - *More than 1 Major Defense KTR* – binary variable: 1 for yes and 0 for no; a program that includes more than one of the contractors listed above = 1
  - *Fixed-Price EMD Contract* – binary variable: 1 for yes and 0 for no

<u>Schedule Characteristics</u>

- RDT&E and Procurement Maturity Measures
  - *Maturity (Funding Yrs complete)* – continuous variable which indicates the total number of years completed for which the program had RDT&E or procurement funding budgeted
  - *Funding YR Total Program Length* – continuous variable which indicates the total number of years for which the program has either RDT&E funding or procurement funding budgeted
  - *Funding Yrs of R&D Completed* – continuous variable which indicates the number of years completed for which the program had RDT&E funding budgeted
  - *Funding Yrs of Prod Completed* – continuous variable which indicates the number of years completed for which the program had procurement funding budgeted
  - *Length of Prod in Funding Yrs* – continuous variable which indicates the number of years for which the program has procurement funding budgeted
  - *Length of R&D in Funding Yrs* – continuous variable which indicates the number of years for which the program has RDT&E funding budgeted
  - *R&D Funding Yr Maturity %* – continuous variable which equals *Funding Yrs of R&D Completed* divided by *Length of R&D in Funding Yrs*
  - *Proc Funding Yr Maturity %* – continuous variable which equals *Funding Yrs of R&D Completed* divided by *Length of Prod in Funding Yrs*
  - *Total Funding Yr Maturity %* – continuous variable which equals *Maturity (Funding Yrs complete)* divided by *Funding YR Total Program Length*
- EMD Maturity Measures
  - *Maturity from MS II in mos* – continuous variable calculated by subtracting the earliest MS II date indicated from the date of the SAR

- *Actual Length of EMD (MS III-MS II in mos)* – continuous variable calculated by subtracting the earliest MS II date from the latest MS III date indicated
- *MS III-based Maturity of EMD %* – continuous variable calculated by dividing *Maturity from MS II in mos* by *Actual Length of EMD (MS III-MS II in mos)*
- *Actual Length of EMD using IOC-MS II in mos* – continuous variable calculated by subtracting the earliest MS II date from the IOC date
- *IOC-based Maturity of EMD %* – continuous variable calculated by dividing *Maturity from MS II in mos* by *Actual Length of EMD using IOC-MS II in mos*
- *Actual Length of EMD using FUE-MS II in mos* – continuous variable calculated by subtracting the earliest MS II date from the FUE date
- *FUE-based Maturity of EMD %* – continuous variable calculated by dividing *Maturity from MS II in mos* by *Actual Length of EMD using FUE-MS II in mos*

- Concurrency Indicators
  - *MS III Complete* – binary variable: 1 for yes and 0 for no
  - *Proc Started based on Funding Yrs* – binary variable: 1 for yes and 0 for no; if procurement funding is budgeted in the year of the SAR or before, then = 1
  - *Proc Funding before MS III* – binary variable: 1 for yes and 0 for no
  - *Concurrency Measure Interval* – continuous variable which measures the amount of testing still occurring during the production phase in months; actual IOT&E completion minus MS IIIA (Jarvaise, 1996:26)
  - *Concurrency Measure %* – continuous variable which measures the percent of testing still occurring during the production phase; (MS IIIA minus actual IOT&E completion) divided by (actual minus planned IOT&E dates) (Jarvaise, 1996:26)

<u>Other Characteristics</u>

- *# Product Variants in this SAR* – continuous variable which indicates the number of versions included in the EMD effort that the current SAR addresses
- *Class – S* – binary variable: 1 for yes and 0 for no; security classification Secret
- *Class – C* – binary variable: 1 for yes and 0 for no; security classification Confidential
- *Class – U* – binary variable: 1 for yes and 0 for no; security classification Unclassified
- *Class at Least S* – binary variable: 1 for yes and 0 for no; security classification is Secret or higher
- *Risk Mitigation* – binary variable: 1 for yes and 0 for no; indicates whether there was a version previous to SAR or significant pre-EMD activities

- *Versions Previous to SAR* – binary variable: 1 for yes and 0 for no; indicates whether there was a significant, relevant effort prior to the DE; a pre-EMD prototype or a previous version of the system would apply
- *Modification* – binary variable: 1 for yes and 0 for no; indicates whether the program is a modification of a previous program
- *Prototype* – binary variable: 1 for yes and 0 for no; indicates whether the program had a prototyping effort
- *Dem/Val Prototype* – binary variable: 1 for yes and 0 for no; indicates whether the prototyping effort occurred in the PDRR phase
- *EMD Prototype* – binary variable: 1 for yes and 0 for no; indicates whether the prototyping effort occurred in the EMD phase
- *Did it have a PE* – binary variable: 1 for yes and 0 for no; indicates whether the program had a Planning Estimate
- *Significant pre-EMD activity immediately prior to current version* – binary variable: 1 for yes and 0 for no; indicates whether the program had activities in the schedule at least six months prior to MSII decision
- *Did it have a MS I* – binary variable: 1 for yes and 0 for no
- *Terminated* – binary variable: 1 for yes and 0 for no; indicates if the program was terminated

The contractor variables in particular require elucidation. The SAR data contains 45 different contractors for the programs in our database. Such a large number of contractors leads to a small number of repeat contractors on different programs, even considering that more than one contractor often work on the same program. These small numbers create problems with coming up with statistically relevant results. Fortunately for our research, the 1990s represented a time of intense defense contractor consolidation. Table 8 shows selected defense contractor consolidations that occur from the period 1993-2000 (Druyun, 2001:4). From these consolidations, we re-categorize our contractors as depicted in Table 9. This gives us sufficient data points for most of the categories to achieve useable results from the regressions.

Table 8. Defense Contractor Consolidations from 1993-2000 (Druyun, 2001:4)

	Boeing	Lockheed Martin	BAE Systems North America	Raytheon	General Dynamics	Northrup Grumman	Litton
Rockwell Defense & Aerospace	X						
Boeing	X						
McDonnell Douglas	X						
Hughes Satellite Systems	X						
General Dynamics Space		X					
GE Aerospace		X					
Martin Marietta		X					
General Dynamics Ft Worth		X					
Lockheed		X					
Loral		X					
IBM Federal Systems		X					
Unisys Defense		X					
LM Sanders			X				
LM Control Systems			X				
Tracor			X				
Convair			X				
General Dynamics Electronics			X				
Viro			X				
Marconi Electronic Systems			X				
Chrysler Defense				X			
E-Systems				X			
Raytheon				X			
Hughes Aircraft				X			
Texas Instruments Electronics				X			
Computing Devices International					X		
Advanced Technology Systems					X		
Lockheed Martin Armament					X		
Bath Iron Works					X		
General Dynamics					X		
NASSCO Holdings					X		
K-C Aviation					X		
GTE Government Systems					X		
Gulfstream Aerospace					X		
Northrop						X	
Grumman						X	
Vought Aircraft						X	
Westinghouse Defense						X	
Logicon						X	
Ryan Aeronautical						X	
TASC							X
Sperry Marine							X
PRC							X
Litton Industries							X
Avondale							X

Regarding the EMD maturity variables, we address both ambiguity and scarcity within the schedule parameters that make up the maturity variables. MS II and MS III dates often have different versions of the same schedule item, making unclear which date to use for computation. For example, a program might have a MS IIA and a MS IIB. The same situation exists for the MS III dates. In order to capture the entire EMD effort, we use the earliest MS II date and the latest MS III date available for our maturity calculations. In our EMD maturity variables that use IOC or FUE for computation, we face a scarcity of data points. In the case of IOC-based maturity computations, 19 of our 90 data points do not have values, which shrinks the database considerably. For FUE-based maturity computations, the database shrinks to only 28 useable data points. The effects of the scarcity of data points somewhat limits the potential use of these as predictors in our regression models.

Table 9. Original Contractors vs. Consolidated Contractors

Original List of Contractor Variables		New List of Contractor Variables
Magnavox	Pratt & Whitney	Lockheed-Martin
McDonnell Douglas	At&T	Northrop Grumman
Bell-Textron	Stewart Stevenson	Boeing
Hughes	Texas Instruments	Raytheon
IBM	Plessey	Litton
GE	E-Systems	General Dynamics
LTV	Motorolla	No Major Defense Contractor
Lockheed-Martin	Avondale	More than 1 Major Defense Contractor
ITT	Bendix	
Westinghouse	Ford Aerospace	
Northrop Grumman	MIDSCO	
Control Data	Honeywell	
Rockwell	Coleman Research	
Boeing	Standard Missile	
United Defense	Loral Voight	
FMC	Osh Kosh	
Sikorsky	Aerojet	
Raytheon	Newport News	
Litton	Teledyne	
EG&G Defense	AIA	
Bechtel	United Technologies	
TRW	GTE	
General Dynamics		

The concurrency indicators allude to the degree to which the production and EMD phases overlap. *Concurrency Measure Interval* and *Concurrency Measure %* we calculate using formulas from RAND's Defense System Cost Performance Database

(Jarvaise, 1996:26). These two variables suffer from the same problem as the FUE and IOC-based maturity computations – using them restricts the number of data points that we can use in our regressions.

Logistic Regression

Logistic regression provides a tool for analyzing possible predictive relationships when the response is either nominal or ordinal. Logistic regression mainly predicts binary outcomes, usually coded '0' and '1' (Neter, 1996:567). In our logistic regression, we seek to develop a model that will predict whether a program will have engineering cost growth or not. Thus, in our historical database, we code a program '1' if it has cost growth and '0' if it has either no cost growth or negative cost growth. We do not concern ourselves with negative cost growth for a pragmatic reason: an estimator would not assess negative cost growth in an estimate. Because we have a distribution of 1's and 0's, we characterize *whether or not a program has engineering cost growth* as a Bernoulli random variable with probability p of success (success=1) (Neter, 1996:568).

Logistic regression takes our historical database of 1's and 0's and estimates the parameters of the model that best fits the predictor values entered into it. Logistic regression is based on the logistic response function and uses the method of maximum likelihood to estimate the parameters that create the best model for the mix of dependent and independent variables (Neter, 1996; Whitehead, 2001). One form of the simple logistic response function is: $\mathrm{Ln}[\,p/(1-p)] = \mathrm{a} + BX + \mathrm{e}$ (Whitehead, 2001:2). This form of the function shows that it essentially represents a linear function with a Y transformation (Whitehead, 2001:2). Applying the natural exponent to both sides of the equation, we isolate [$p/(1-p)$] and acquire the following equation: [$p/(1-p)$] = $\exp^{a} \exp^{B}$ $\exp^{X} \exp^{e}$. The left side of the equation is called the odds ratio. This ratio represents the probability of success (1) divided by the probability of failure (0).

Dr. Whitehead of East Carolina University discusses the usefulness of the odds ratio and the interpretation of the coefficient B in "An Introduction to Logistic Regression." He mentions that one cannot interpret the coefficient of the independent variable X in the logistic function the same way that one would for a linear regression. One can gain an understanding of the effect of the coefficient in logistic regression by considering the effect on the odds ratio in the one-variable model (this interpretation does not apply to multiple-variable models). As X increases by one unit, the odds ratio increases $\exp^{B}$. In our situation, where '1' = cost growth and '0' = no cost growth, if $\exp^{B}=3$, then as X increases by one unit, our chance of experiencing cost growth increases three-fold (Whitehead, 2001:2-3).

Neter et al. add to the description of the logistic response function coefficients by describing the graph of the function. "A logistic response function is either monotonic increasing or monotonic decreasing, depending on the sign of B_1. Further, it is almost linear in the range where E{Y} is between .2 and .8 and gradually approaches 0 and 1 at the two ends of the X range" (Neter, 1996:571). The version of the formula Neter et al. use to plot the function follows: $E\{Y\} = \exp^{(B_0+B_1X)} / (1+\exp^{(B_0+B_1X)})$. In the terms Dr. Whitehead uses, the equation is: $p = [\exp^{(a+BX)}] / [1+\exp^{(a+BX)}]$ (Whitehead, 2001:2). Figure 4 shows the reason for the use of the logistic regression: the function constrains itself to values between zero and one. This particular figure shows the value of E{Y} (or *p*) when *B* is positive. As *X* increases, *p* increases. When *B* is negative, as *X* increases, *p* decreases (Neter, 1996:571).

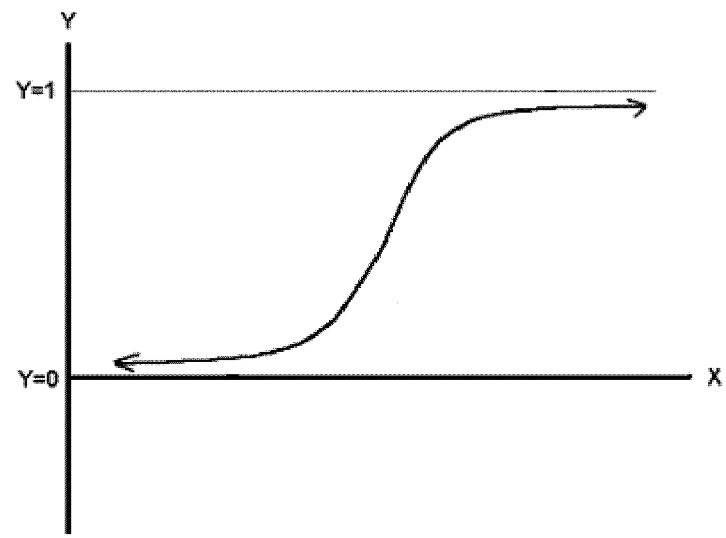


Figure 4. Logistic Regression Function (Whitehead, 2001:3)

We use JMP® 4 (SAS Institute, 2001) software to accomplish the logistic regression in order to help us identify the best model for estimating whether or not a program will have cost growth. JMP® uses maximum likelihood to estimate the coefficients of our model. Because JMP® has no automatic method, such as stepwise, for logistic regression, we manually compute thousands of individual regressions, recording our results on spreadsheets. We start with one-predictor models of all possible variables. Then we regress using all combinations of two-predictor models and record the results. We continue this process, eventually whittling down the best combinations for use at the next level in order to cut down on the amount of regressions necessary. We stop when we reach a model for which the gain of adding another variable does not warrant the additional complexity of the model that another variable adds. We intend to find several candidate models for each number of predictors and then narrow down to the best one for each number of predictors and validate the model using about 20 percent of the data that we set aside for validation.

Multiple Regression

In order to discover prediction models for the percent of engineering cost growth based on more than one predictor variable, we use multiple regression. As with logistic regression, we use JMP® for the multiple regression analysis. We use the stepwise method to identify those predictor variables that have a statistically significant impact on the ability of the model to predict our response variable, *Engineering %*. From our stepwise analysis, we build models using the standard least squares method, whereby JMP® estimates the form of the functional relationship between the predictors and the response variable that minimize the sum of squared deviations from the predicted values at each level of the predictors (Neter, 1996).

Because of the large amount of candidate predictor variables, we exceed JMP®'s stepwise calculation abilities when we include all of our variables in a single run. In addition, we seek models with varying numbers of predictors. Thus, we must repeat the stepwise and standard least squares several times in order to achieve the desired results. As with logistic regression, we discover several candidate models for each number of predictors. Then we narrow our results to the best model for each number of predictors. We continue adding variables to the model until the number of variables equals about one tenth of the number of data points used in the model; this ensures we do not over-fit the model (Neter, 1996:437). We check the model's robustness using the same validation data as for the logistic regression.

We build four regression models that we briefly introduce in this paragraph. We build one logistic model using 90 data points. This model predicts whether a program will have engineering cost growth in RDT&E dollars. To simplify our analysis, we call this Model A. We then build three multiple regression models. We call Model B the model that we build from the 47 of the 90 data points that do have cost growth. We apply a log transformation to the response variable in this model to correct for heteroskedasticity in the residual plot. We build Model C as an alternative to Model B. Model C is the same as Model B except that we do not transform the response variable. Model D represents what would happen if we skip logistic regression and use stepwise and multiple regression on all 90 data points (ignoring the problems of heteroskedasticity in the residuals, and ignoring the fact that we do not desire to predict negative cost growth).

Chapter Summary

This chapter sets forth our analytical process. In it we demonstrate the tie between the literature review and the analysis we perform. Further, we explore the credibility of the SAR data, describe the process by which we compile the data into a useable spreadsheet format, and describe the predictor variables that we will investigate in our models. Finally we explain the reasoning for our use of logistic and multiple regression techniques and the process into which we incorporate these techniques.

IV. Results and Discussion

Chapter Overview

This chapter explicates the results of both the logistic and multiple regression analysis. In it we describe the resulting models and their robustness. We also analyze the models for statistical validity and practical usefulness. We evaluate all four families of models (A, B, C, and D) for each number of predictor variables we use. We name the resulting models after the family and number of variables we use. For example, A.1 refers to the logistic regression model that uses only one predictor variable, and B.3 refers to the multiple regression model that has three predictor variables using data from only those programs that have cost growth and for which we perform a natural log transformation on the response variable.

Preliminary Data Analysis

Initially, we set out to produce only one model that will predict the amount of engineering cost growth a program will incur given certain program characteristics. To do this, we plan to use multiple regression. However, a look at the distribution of the response variable *Engineering %* via Figure 5 reveals a mixed distribution: it has a discrete mass at zero and a continuous distribution elsewhere. As we discuss in Chapter III, this propels us to explore the possibility that a two-step model might produce superior results. Thus, we formulate Model A for use in determining whether a program will have cost growth or not, followed by Model B to determine how much cost growth will occur if Model A indicates cost growth will occur. Model C is an option that we compare to

Model B, and we create Model D as an option to using either A alone (in the case that it predicts no cost growth), A and B together, or A and C together. We start our analysis with Model A and work our way alphabetically to Model D.

Stem	Leaf	Count
1	7	1
1		
1	3	1
1	11	2
0		
0	6	1
0	4555	4
0	22222222223333	14
0	1111111111	10
-0	111100000000000000000000000000000000000000000000000000	54
-0	33	2
-0		
-0		
-0		
-1		
-1	3	1
-1		

Figure 5. Stem and Leaf Plot of *Y* (*Engineering %*, stem in 100's, leaf in 10's)

Logistic Regression Results – Model A

As mentioned in Chapter III, no stepwise-type function exists in JMP® for logistic regression. Without this automated procedure to narrow down our predictors, we face an enormous number of possible combinations of variables to research. In fact, using our 78 variables to explore all possible combinations of one through seven-variable models requires over 2.6 billion independent regressions. Given the enormity of exploring all of these combinations, we narrow our predictor combinations to only those that show the most promise as we progress from a single-variable model to a seven-variable model.

The process we use to narrow predictor variables deserves some attention in order to give the reader an appreciation for its copiousness. We begin by regressing all one-variable models and recording the results. We select the best nine, one-variable models and regress all possible two-variable models that stem from each of those nine, one-variable models. We then select the eight models that stand out among the two-variable results and regress all possible three-variable models that stem from each of those eight, two-variable models. We continue this process until the benefit of adding variables does not outweigh the complexity of the resulting model.

For each number of variables we try, we have anywhere from seven to twelve models that carry forward for regression with additional variables. This process culminates in approximately four thousand regressions and seven generations of models – one generation for each number of predictors. Within each of these seven generations, we then compare the several candidate models and select the best model. Table 10 summarizes the statistical characteristics of the resulting models. We select these models over other candidate models based on the measures listed in the table. The following paragraphs discuss these measures.

Table 10. Evaluation Measures for Model A

Evaluation Measures	Number of Predictors						
	1	2	3	4	5	6	7
R^2 (U)	0.1577	0.2178	0.2856	0.3256	0.3660	0.5050	0.6012
Number of Data Points	87	87	75	75	75	61	61
Area Under ROC Curve	0.7678	0.7906	0.8293	0.8542	0.8659	0.9264	0.9481

First we compare models based on R^2 (U). This measure of fit differs in its interpretation of R^2 from linear models. David Garson in his online textbook explains the difference:

> There is no widely-accepted direct analog to OLS [ordinary least squares] regression's R^2. This is because an R^2 measure seeks to make a statement about the "percent of variance explained," but the variance of a dichotomous or categorical dependent variable depends on the frequency distribution of that variable. For a dichotomous dependent variable, for instance, variance is at a maximum for a 50-50 split and the more lopsided the split, the lower the variance. This means that R-squared measures for logistic regressions with differing marginal distributions of their respective dependent variables cannot be compared directly, and comparison of logistic R-squared measures with R^2 from OLS regression is also problematic. Nonetheless, a number of logistic R-squared measures have been proposed. (Garson, 2002:9)

Garson goes on to describe several alternative measures that give a measure comparable to the R^2 of OLS regression, but he mentions that these measures "are not goodness-of-fit tests but rather attempt to measure strength of association." The R^2 (U) that JMP® uses is the difference of the negative log likelihood of the fitted model minus the negative log likelihood of the reduced model divided by the negative log likelihood of the reduced model. As with the traditional R^2, a higher R^2 (U) indicates a better model. The JMP® help menu says about its R^2 (U), "high R^2 (U)s are unusual in categorical models" (JMP®, 2001: Help). Thus, we look for a high R^2 (U) but temper our expectations in light of this comment and understand the interpretation of the R^2 (U) differs from that of OLS R^2. The models we select all have the highest R^2 (U)s of any of the other models within the same generation of predictors.

Next, we consider the number of data points. The number of data points plays a particularly important role, because the higher the number of data points, the more of our population we capture in our sample. Thus, our sample becomes more representative of the population. In addition, the larger the sample size, the more predictor variables we can add before the model becomes invalid statistically. According to Neter et al., a model should have at least six to ten data points for every predictor used. Thus, in this study, if a model falls below ten data points per predictor, then we carefully consider the additional benefits to the model gained by adding the variable. Any model in which the ratio of data points to predictors falls below six we eliminate as a possibility (Neter, 1996:437). The seven-variable model has only 8.7 data points per predictor, all the rest have over ten data points per predictor. Thus, we carefully weigh the additional benefit of the seventh variable in the model when selecting the best model, and we negate the possibility of an eight-variable model.

Third, we consider the *p*-value associated with the Chi-squared statistic for the whole-model test. Garson describes this statistic as follows:

> Model chi-square provides the usual significance test for a logistic model. Model chi-square tests the null hypothesis that none of the independents are linearly related to the log odds of the dependent. That is, model chi-square tests the null hypothesis that all population logistic regression coefficients except the constant are zero. It is thus an overall model test which does not assure that every independent is significant. (Garson, 2002:8)

Hence, we use this measure for the same purpose as for OLS regression – to test whether the model as a whole predicts significantly better than the reduced model. A *p*-value less than 0.05 tells us the model has statistical significance as a predictive model. Because all of the logistic regressions have *p*-values less than 0.0001, this measure does not help us discriminate between models. Thus, we do not include *p*-values in the table.

The last whole-model measurement we consider is the area under the receiver operating characteristic (ROC) curve. The medical field routinely uses logistic regression, and in particular ROC curves, so we look to their experts for insight into the measure. Clifford S. Goodman of the Lewin Group (a medical consulting firm) provides an interpretation of the ROC curve. The curve itself maps out the proportion of the true positives (sensitivity) out of all actual positives versus the proportion of false positives (1-specificity) out of actual negatives, both calculated across all possible calibrations of the model.

In our experiment, we define a true positive as a program for which the model correctly predicts that cost growth will occur in the fitted values. For a false positive, the model incorrectly predicts that cost growth will occur in the fitted values. The calibrations represent the cutoff probabilities that differentiate between whether a program receives a one or a zero in the logistic regression. The area under the ROC curve, then, gives an idea of the probability associated with ability of the model to accurately predict whether a program will have cost growth, based on results from the fitted values (Goodman, 1998:Appendix A). Of all the measures, this one has the most pertinence, since it deals most specifically with our goal of accurately assessing whether a program will or will not have cost growth. As with the other whole-model measures, we find that the measure improves as we add more predictor variables through the addition of seven predictors.

Table 11 displays the *p*-values for the parameter estimates. Just as in OLS regression, a lower *p*-value indicates higher statistical significance for that parameter as an estimator of the response variable. A good model should have *p*-values less than 0.05.

In fact, we desire the *p*-values as low as possible in order to hedge against over-fitting the model (tailoring the model to the fitted data to the extent that it lessens the ability of the model to predict the response values of the population). Only the five-variable model in Table 11 breaches the 0.05 criterion. Because *Length of Prod in Funding Yrs* is borderline significant (0.0507), we do not disqualify this variable as a candidate estimator. Thus, we consider all the models listed in Table 11 as potential candidates for modeling whether a program will have cost growth.

While Table 10 and Table 11 demonstrate how models fare individually against the measurement criteria, selecting a best model requires some means of comparison among the different levels of predictors. In order to visualize the combined impact that the incremental addition of predictors has on the various measures of effectiveness for the logistic model, we create Table 12. Specifically, this table shows the increase or decrease in each evaluation measure as we add a single predictor to a given model. For example, as we add a predictor to the model with one independent variable, we gain 0.0601 in R^2 (U) and our ratio of data points to the number of independent variables in the model decreases to 43.5.

Table 11. *P*-Values of Predictor Variables for Model A

Predictor Variables	Number of Predictors						
	1	2	3	4	5	6	7
Maturity from MSII (in mos)	0.0002	0.0076					
Length of R&D in Funding Yrs		0.0100		0.0288	0.0059	0.0015	0.0020
RAND Modification			0.0091	0.0043	0.0021	0.0022	0.0037
Actual Length of EMD (MSIII-MSII in mos)			0.0175	0.0039	0.0041	0.003	0.0029
Funding Yrs of R&D Completed			0.0006				
MSIII-based Maturity of EMD %				0.0187	0.0219	0.0202	0.0148
Length of Prod in Funding Yrs					0.0507	0.0031	0.0012
Actual Length of EMD (using IOC-MSII in mos)						0.0334	0.0154
Land Vehicle							0.0132

Table 12. Incremental Changes in Evaluation Measures for Model A

Evaluation Measures	Number of Predictors						
	1	2	3	4	5	6	7
Incremental increase in R^2 (U) with additional predictor	0.1577	0.0601	0.0678	0.0400	0.0404	0.1390	0.0962
Ratio of data points to number of variables	87.0	43.5	25.0	18.8	15.0	10.2	8.7
Incremental increase in Area Under ROC Curve with additional predictor	0.2678	0.0228	0.0387	0.0249	0.0117	0.0605	0.0216

From Table 12, we create Figure 6 to better observe the effects of the marginal change in the number of predictors. Figure 6 shows the changes on the whole-model measures with each one-predictor increase. In this graph, the higher numbers indicate that the addition of the extra predictor affects a more significant impact than that of a lower number.

From the graph, we see similarities in the behavior of the whole-model measures. The addition of the first predictor and the sixth predictor show the greatest increases for area under the ROC curve and R^2 (U). Both measures have relatively low marginal gains at the addition of predictors four and five.

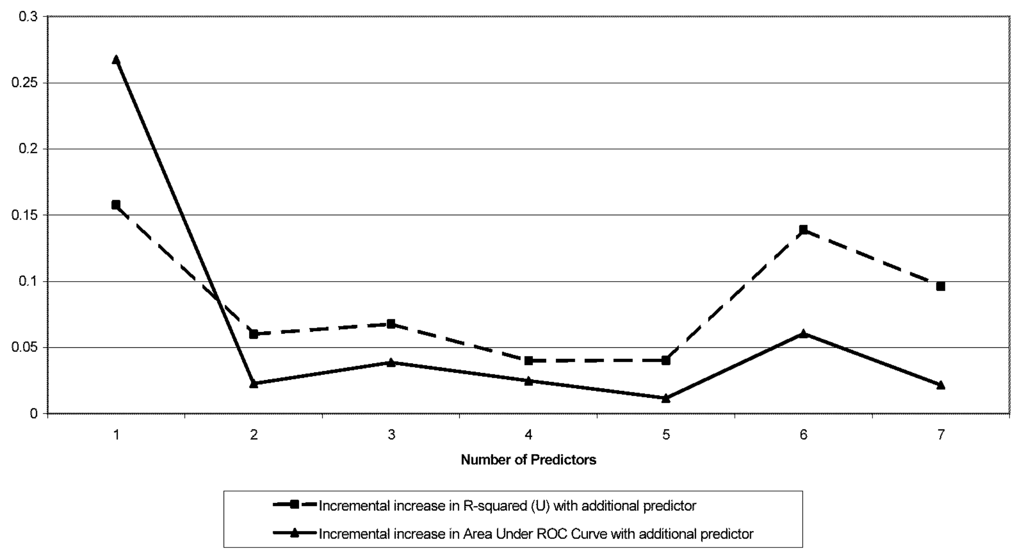


Figure 6. Incremental Changes in Whole-model Measures for Model A

We view with particular interest the spike in the measures at variable six and the drop-off at the addition of variable seven. Specifically, we see an increase in the R^2 (U) of 0.139, which increases significantly the association between the six-variable model and the outcomes. Secondly, this model increases the probability under the ROC curve by 0.06048 to 0.92641, making the probability of capturing all true positives high and the probability of having false positives low. Therefore, the gains of adding the sixth variable outweigh the complication of the model by adding the sixth variable. Because we desire to maximize our ability to correctly predict whether a program will have cost growth, we consider whether the seven-variable model might satisfy our needs without over-fitting the model.

Upon first glance, the drop in marginal return for the addition of the seventh variable seems an indication that such a model over-fits the data. In addition, earlier in this text we convey concerns about the ratio of data points to independent variables. On the other hand, the amount of increase in R^2 (U) approaches 0.1, a measurable increase, and it accounts for the possible cost growth associated with a land vehicle. For these reasons, we preliminarily consider the seven-variable model as the best model, based on the whole-model measures (Appendix A). Validation of the models will show whether this conclusion will perdure.

For validation, we use 25 data points that we randomly select from the original 115-point data set. Of these 25 data points, 12 data points have missing values for some of the variables, leaving 13 for validation. These 13 data points represent approximately 17.6 percent of the 61 viable data points the model uses. Although we fail to meet our goal of validating using 20 percent of the data, we are relatively close to this goal. Thus, we have a reasonable degree of confidence in the results.

The validation process entails saving the functionally predicted values ('0' or '1') in JMP® for each of the validation data points and comparing those values to the actual values. JMP® computes the predicted values by assessing the probability of having cost growth. JMP® assigns a '1' to any point with a probability of 0.5 or greater and a '0' otherwise. The user can change these defaults to make the model more or less conservative, but in our case, we use the default setting of 0.5. Upon validation, the model accurately predicts nine out of the 13 data points for a success rate of 69 percent, further evidencing that this model has some predictive ability, and establishing it as our best model (Appendix A).

Multiple Regression Results – Model B

We build model B for those situations where a decision maker knows that a program will have cost growth and wants to know the amount of expected cost growth the program. To build this model, we start with our randomly selected 90 data points and exclude programs that have no cost growth, leaving us with 47 data points. Using only these points should give the model more accuracy to predict, since it prevents data points outside the range of interest from skewing the results. We use the same pool of candidate predictor variables as in Model A, and for the *Y* variable we use *Engineering %*, which measures the percent increase of engineering cost growth from the DE.

Upon a preliminary analysis of the data, we notice the *Y* variable does not have a normal distribution (Figure 7). In fact, *Y* exhibits more of a lognormal distribution. Running a few test regressions reveals that strong patterns exist in the residual plots (Figure 7). The plots fail the Breusch-Pagan test (for constancy of variance) by large margins (Neter, 1996:115). Based on these findings, we perform a natural log transformation of the *Y* variable. This transformation successfully dispels the heteroskedasticity previously found (Figure 8). The transformation also results in a distribution shape much closer to normal, though still slightly skewed right. The Shapiro-Wilk test indicates the normal distribution sufficiently fits the data at an alpha of 0.05 (Figure 8). We use this natural log transformation for all Model B regressions.

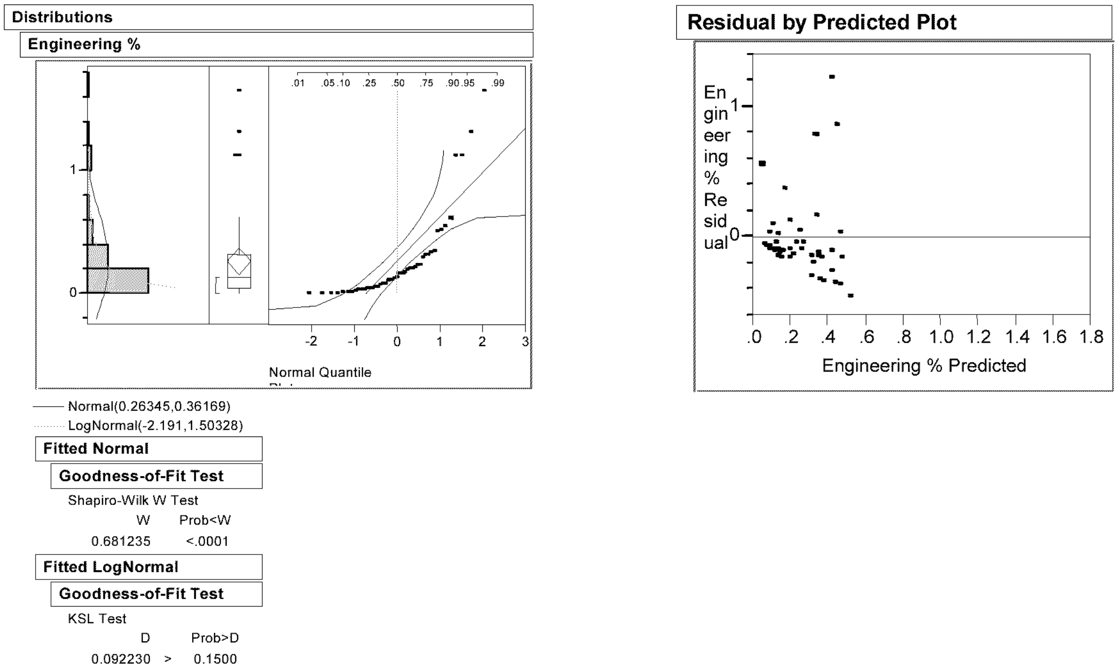


Figure 7. Distribution of *Y* and Residual Plot of Untransformed Model B

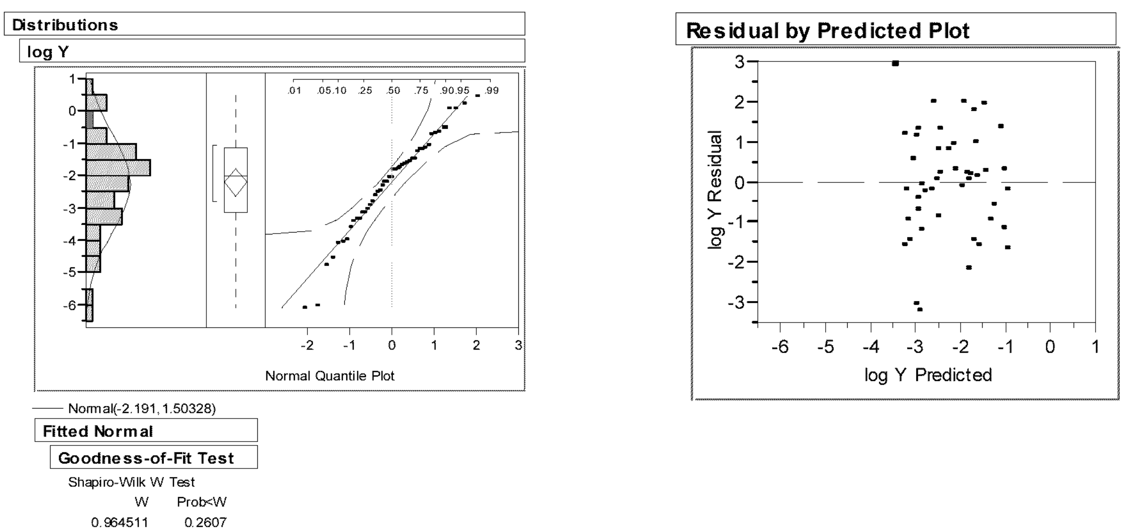


Figure 8. Distribution of Log *Y* and Residual Plot of Transformed Model B

Stepwise regression helps us narrow the predictor variables. Since we start with only 47 data points, we limit the number of predictors to five in order to keep the predictor to data point ratio from going too far below ten to one (Neter et al., 1996:437). We produce several regression models for each number of predictors, just as we do for Model A. We then choose the model that provides the best predictability while maintaining statistical significance as a model. We summarize the results of the best regressions for each generation of variables in Table 13 and Table 14.

Table 13. Evaluation Measures for Model B

Evaluation Measures	Number of Predictors				
	1	2	3	4	5
R^2 Adj	0.2200	0.3386	0.4645	0.4743	0.4934
Number of Data Points	46	46	42	42	43
P-Value ANOVA	0.0006	0.0001	0.0001	0.0001	0.0001

Table 14. *P*-Values of Predictor Variables for Model B

Predictor Variables	Number of Predictors				
	1	2	3	4	5
Maturity from MSII (in mos)	0.0006	0.0018	0.0069	0.0015	
No Maj Def KTR		0.0047	0.0024	0.0004	0.0068
PAUC			0.0410	0.0069	0.0004
Class At least S				0.0355	
Svs>1					0.0273
R&D Funding Yr Maturity %					0.0029
Total Funding Yr Maturity %					0.0024

We find all of these models comply with the underlying assumptions of constant variance and normality for linear regression at an alpha of 0.05. We assume independence for no obvious serial correlation is present, and we have removed dependent programs in the data set. In addition, we test the predictors for multicollinearity by ensuring that all variance inflation factors (VIFs) as calculated by JMP® are less than ten (Neter, 1996:387).

From Table 13, we notice a few general patterns in the data. First, as the number of variables increases, the adjusted R^2 increases. This indicates that the model explains more of the variance as we add variables up to five. Also, the number of viable data points decreases to 42 when we add the third variable, but it does not decrease thereafter. The emolument in predictive power gained by adding that third variable warrants such an addition, and for variables four and five adjusted R^2 increases free from trade-off in data. In fact, the five-variable model adds a data point. The analysis of variance (ANOVA) *p*-value remains constant for all generations of predictors save the first; thus, this measure does not help us discriminate. A look at the significance levels of the predictor variables in Table 14 shows that all predictors significantly add to the model at an alpha level of 0.05. The least significant of the predictors, *PAUC*, occurs in Model B.3, with a significance of 0.0410. As with Model A, we chart the changes in these measures (Table 15).

Table 15. Incremental Changes in Evaluation Measures for Model B

Evaluation Measures	Number of Predictors				
	1	2	3	4	5
Incremental increase in R^2 Adj with additional predictor	0.2200	0.1186	0.1259	0.0098	0.0191
Ratio of data points to number of variables	46.0	23.0	14.0	10.5	8.6

From Table 15, we see the largest marginal increase in adjusted R^2 at variable one and the smallest at variable four. A fourth variable increases adjusted R^2 by less than

0.01. This modest increase in adjusted R^2 does not call for the addition of a fourth variable. Thus, Model B.3 represents the model where the costs of adding more predictor variables exceed the predictive benefits according to the measures we have before validation.

For validation, we use the same data as for Model A. Only 14 out of the original 25 validation data points have cost growth; the other 11 do not. The 14 represent roughly 25 percent of the overall data used to create the model plus the validation points, giving us enough points to result in a credible validation. During model validation, the first two models use all 14 data points, while the last three only use 13 because of missing data for some of the predictor variables.

For validation of the range estimates, we originally consider 95 percent prediction intervals (PIs). However, after back-transforming the *Y* via the natural exponential function, we find these PIs impractically wide in some cases. In order to compensate somewhat for the wide PIs, we use an 80 percent PI. We believe this smaller interval will prove more useful to a user. For an 80 percent interval, we expect to see about 80 percent of the validation data points fall within it. For the models that use less data points (usually those with a higher number of variables in the model), we expect to see fewer data points fall within the PIs because of the increased variability associated with smaller sample sizes. Table 16 displays the results of our validation.

Table 16. Validation Measures for Model B

Validation Measures	Number of Variables				
	1	2	3	4	5
Obs Within 80% PI	78.57%	78.57%	69.23%	69.23%	61.54%
Avg Width of PI (Eng %)	59.59%	82.37%	75.67%	86.29%	61.88%
Obs Below 90% UB	100.00%	100.00%	92.31%	92.31%	84.62%
Obs Above 90% LB	78.57%	78.57%	76.92%	76.92%	76.92%
Mean Absolute Deviation	18.88%	17.23%	18.24%	19.23%	19.01%

The first four measures in this table assess the appropriateness of the model for the validation data, while the last measure assesses the appropriateness of the model for both the data used to build the model and the validation data. The first two measures tell us the percent of observations that fall within the 80 percent PI and the average width of the PI respectively. The next two measures relate closely to the first two. The first of these assesses the percent of observations that fall below a 90 percent upper bound (UB), and the other measure assesses the percent of observations that fall above a 90 percent lower bound (LB).

From these measures we discover the data points tend to violate the lower bound more than the upper bound. That is, we empirically expect to see 90 percent of the data fall in both categories. For the UB, the model meets this expectation. For the LB, it is not. Considering the small validation sample size and the skewed right property of a lognormal distribution, this trend is not unexpected and not a source of concern.

With respect to usefulness for a cost estimator, we investigate the average PI widths and mean absolute deviation. The average PI widths, measured in engineering cost growth as a percent of the DE, vary from the low 60's to the high 80's. This represents a considerable spread, and highlights the variability still present in modeling *Engineering %*. This variability coupled with small validation sample size suggests this descriptive measure has limited usefulness as a comparison tool.

The final measure in Table 16, the mean absolute deviation, assesses the accuracy of the point estimate. We calculate it using the formula, $\sum_{i=1}^{n} \left| predicted(i) - actual(i) \right| \Big/ n$, for all 115 data points. We measure the mean absolute deviation in percent engineering cost growth, so interpretation proves straightforward. The lower the mean absolute deviation, the better the model's predicted values fit the entire data set. Mean absolute deviation gives a measure to compare with adjusted R^2 to see how the models fit with validation data versus without.

In general, the two-variable model has a slightly better mean absolute deviation than the three-variable model. However, this difference does not induce us to overturn our initial assessment of Model B.3 as the best model in terms of adjusted R^2. Thus, although all five models perform reasonably well in predicting the percent of engineering cost growth, Model B.3 performs most efficaciously (see Appendix B for model).

Multiple Regression Results – Model C

As demonstrated in the previous section, Model B performs fairly well as a predictive formula. In order to compare Model B to a more simplistic regression approach, which we show later is an incorrect methodology, we attempt to regress using a model with a non-transformed *Y*, which we call Model C. We use stepwise regression to narrow the predictors, and then we use OLS regression to build our models, just as in

Model B. All conditions of the regression procedure remain the same for C as for B, with the exception of the *Y* transformation.

We attempt several models for each number of variables. None of the models we attempt passes the Shapiro-Wilk test for normality of residuals, and none of the models passes the Breush-Pagan test for constancy of variance (both at an alpha of 0.05). In addition, in almost every model we attempt, an influential outlier exists, which is defined as having a Cook's Distance greater than 0.5 (Neter, 1996:381). Most of the time removing the outlier leads to several other influential outliers. Thus, we could not avoid violations of the basic principles that underlie OLS regression; Figure 9 shows an example of such violations.

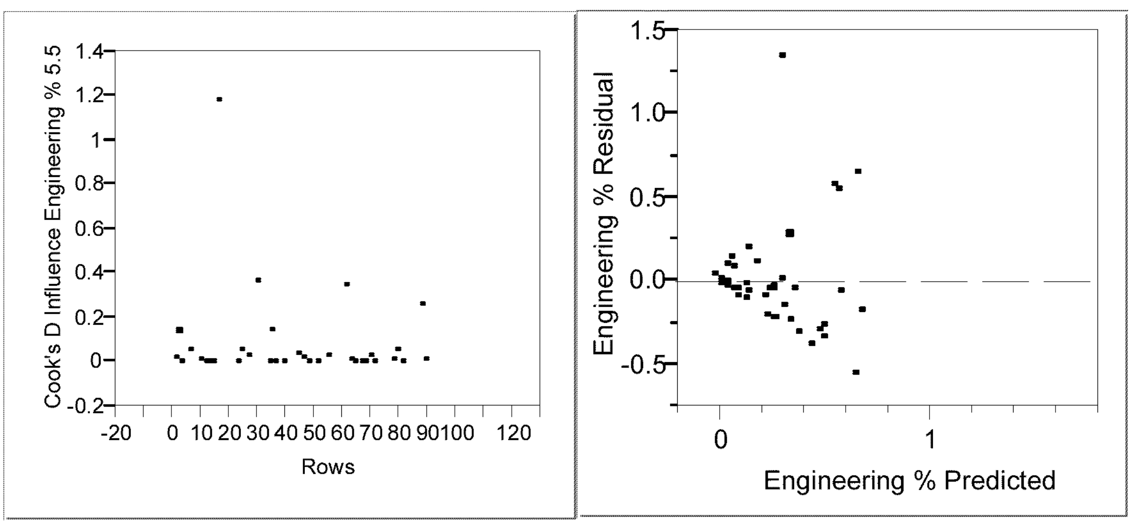


Figure 9. Cook's Distance and Residual Plot of Model C

Table 17 contains the best model for each number of predictors. Within these results, all models violate normality and constancy of variance, and Models C.1, C.4, and C.5 contain influential outliers that we cannot neutralize by exclusion of the influential data point. Although regression and ANOVA are robust techniques for violations of normality and constant variance, our inferential diagnostics, such as *p*-values and prediction intervals may be invalid. Consequently, caution is warranted. Model C.2 originally has an influential outlier, but by removing the Bradley Fighting Vehicle M2/M3 we satisfy the Cook's Distance threshold with a measure of 0.46. Model C.3 also scarcely meets the influential outlier threshold with a Cook's Distances of 0.5.

Table 17. Evaluation Measures for Model C

Evaluation Measures	Number of Predictors				
	1	2	3	4	5
R^2 Adj	0.1924	0.3120	0.3000	0.3988	0.4365
Number of Data Points	41	47	38	38	38
P-Value ANOVA	0.0024	0.0001	0.0016	0.0003	0.0002

The adjusted R^2 measure varies between 0.19 and 0.44. Interestingly, the adjusted R^2 decreases slightly with the addition of the third variable. All the models use a good portion of the available data points. The five-variable model uses 38 data points, giving it a ratio of about 7.6 data points per variable. This ratio approaches the limits of adequacy for a data point to variable ratio, nevertheless, we keep the model for consideration.

Table 18 shows that the specific predictor variables Model C uses vary considerably with the number of predictors. From the table, we also notice a wide spread in the *p*-values of the predictors. A cursory examination of the *p*-values reveals that the four-variable model has a predictor with questionable significance at an alpha level of 0.05, and Model C.5 has two with questionable significance at that alpha. Moreover, Model C.5 has one variable with a significance level well above 0.05 and approaching 0.10. The models have comparable ANOVA *p*-values, with C.1 being slightly larger. Thus, we use adjusted R^2 to discriminate among the models.

Table 18. *P*-Values of Predictor Variables for Model C

Predictors	Number of Predictors				
	1	2	3	4	5
IOC-Based Maturity of EMD %	0.0024				
No Maj Def KTR		0.0075	0.0153	0.0362	0.0027
Funding Yrs Prod Completed		0.0026			0.0006
Maturity from MSII in mos			0.0180		
Actual Length of EMD MSII-MSIII in mos			0.0190		0.0514
MSIII-based Maturity %				0.0012	
Air				0.0136	
Land					0.0946
Class at Least S					0.0462
Versions Previous to SAR				0.0547	

Table 19 shows a considerable rise in adjusted R^2 at the addition of variables one, two, and four (also see Figure 10). Variable three induces a very slight fall in adjusted R^2, while the addition of the fifth variable brings a relatively small increase in adjusted R^2 which does not outweigh the worsening of the data point to variable ratio which it causes. From this information, we eliminate the five-variable model. We also eliminate the one-variable model, because the gains in adjusted R^2 of the two-variable model warrant superceding C.1. The predictor, *Versions Previous to SAR*, in Model C.4 barely surpass significance at alpha of 0.05. Despite this slight breach in significance, we select Model

C.4 as the most promising pre-validation model, because the increase of adjusted R^2 from approximately 0.31 to 0.40 improves the model's predictive ability.

Table 19. Incremental Changes in Evaluation Measures Model C

Evaluation Measures	Number of Predictors				
	1	2	3	4	5
Incremental increase in R^2 Adj with additional predictor	0.1924	0.1196	-0.0120	0.0988	0.0378
Ratio of data points to number of variables	41.0	23.5	12.7	9.5	7.6

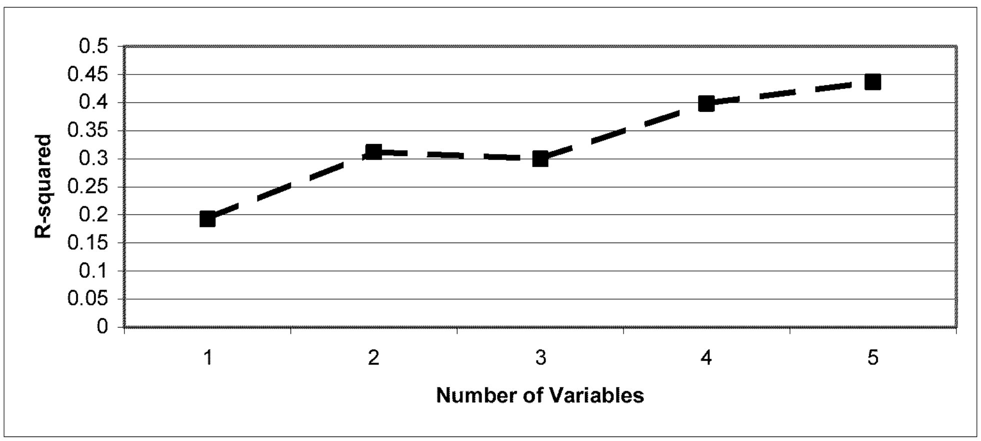


Figure 10. Changes in Adjusted R^2 for Model C

We use the same procedures and data to validate Model C as those for Model B. Validation of the model yields the results shown in Table 20. The models all do fairly well at range prediction of the validation data: on average about 79 percent of the validation data lies within the intervals of each model. The average widths of the intervals range from 64 to 92 percent, again problematically wide, and the models seem to predict below the upper bound and above the lower bound equally well. Although Model C.4 fares the worst in this measure, all models fare similarly, and the difference does not have a magnitude such that by itself it changes our assessment of C.4 as the best model.

Table 20. Validation Measures for Model C

Validation Measures	Number of Variables				
	1	2	3	4	5
Obs Within 80% PI	92.31%	64.29%	81.82%	72.73%	81.82%
Avg Width of PI (Eng %)	91.78%	64.40%	84.03%	78.61%	77.51%
Obs Below 90% UB	92.31%	85.71%	90.91%	81.82%	90.91%
Obs Above 90% LB	100.00%	78.57%	90.91%	90.91%	90.91%
Mean Absolute Deviation	22.55%	21.81%	24.11%	25.70%	23.74%

Comparison of Models B and C

Figure 11 compares the adjusted R^2 at each predictor level. For Models B and C, for all levels of predictors, Model B outperforms Model C in this measure. Model B.3 even exceeds the predictive capabilities of the four and five-variable versions of Model C. A comparison of the mean absolute deviations yields similar results: Model B's mean absolute deviations are smaller at all levels of the predictors (Figure 12). Unlike adjusted R^2, the mean absolute deviation takes into account the validation points in its assessment of fit. Thus, this measure tends to give a better idea of the population fit of the model, since it includes more of the population. For both measures of point estimation accuracy, Model B's performance exceeds that of Model C for each level of predictor.

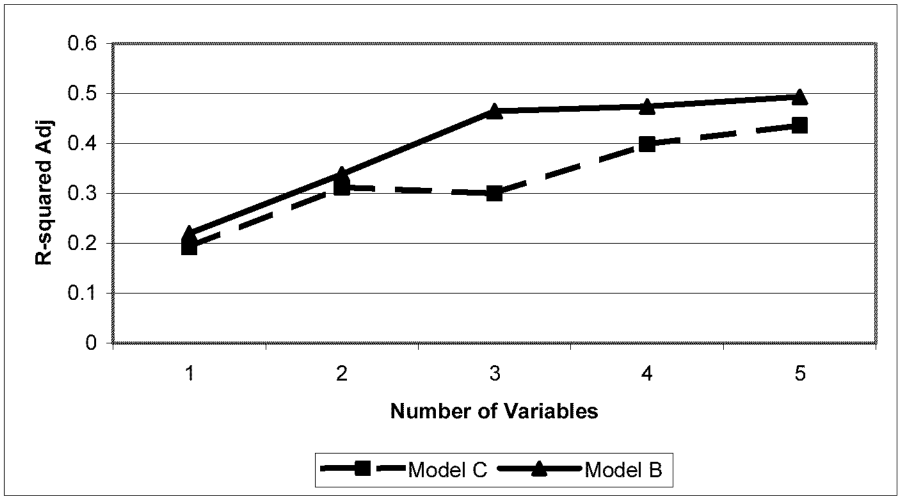


Figure 11. Comparison of Adjusted R^2 for Models B and C

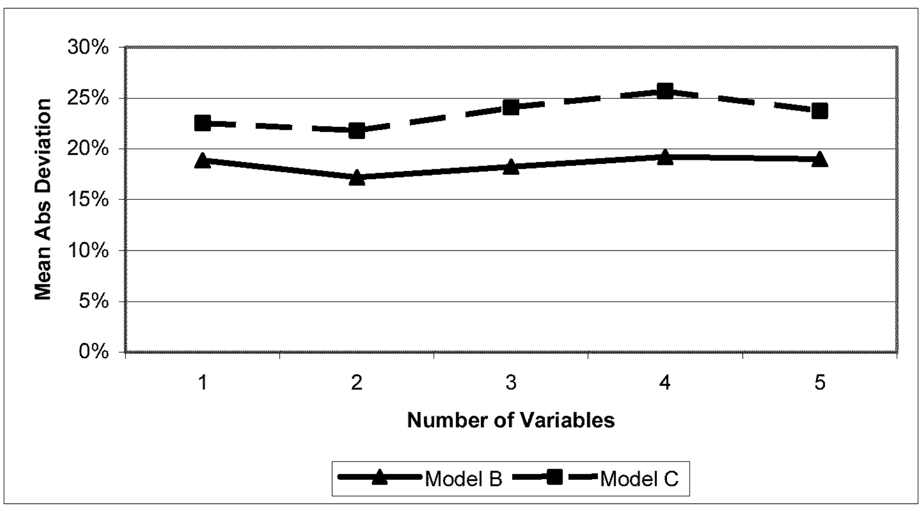


Figure 12. Comparison of Mean Absolute Deviation for Models B and C

The lack of consistency of predictor variables in Model C contrasts with the more consistent Model B. *No Maj Def KTR* appears in Models C.2 through C.5, and *Funding Yrs Prod Completed* and *Actual Length of EMD MS III-MS II in mos* repeat once in the family of C models. Beyond these variables, no consistency exists. Model B, however, repeats the use of *Maturity from MS II (in mos)* and *No Maj Def KTR* four times each. This model also uses *PAUC* three times. Thus, the results of the parameter selection of Model C appear somewhat more erratic than those of Model B.

For interval estimation, we rely on the validation results of the two families of models (Table 16 and Table 20). We notice first that the mean of the average widths of the PIs for Model B equals 74.84 percent, while the mean of the average widths of the PIs for Model C equals 79.27 percent. These measures do not differ much, and the difficulties associated with comparing the prediction intervals (i.e., lack of customary OLS assumptions with Model C) lead us to assign to this observation only minor influence in our decision-making.

We do not expect the non-transformed model to possess comparable range estimates, because of the violations of the normality and constant variance assumptions of OLS regression. Surprisingly, Model C seemingly performs on par with Model B during validation (Figure 13). However, the PI that Model C produces does not represent a true 80 percent PI, because of the violation of the assumptions of the OLS model. Although we witness that Model C's PIs do well at capturing the validation data, we do not really know in fact what prediction level the interval represents. This inferential uncertainty coupled with the results from Figure 13 and Figure 14 leads us to support the use of Model B over Model C as both a point estimator and a range estimator.

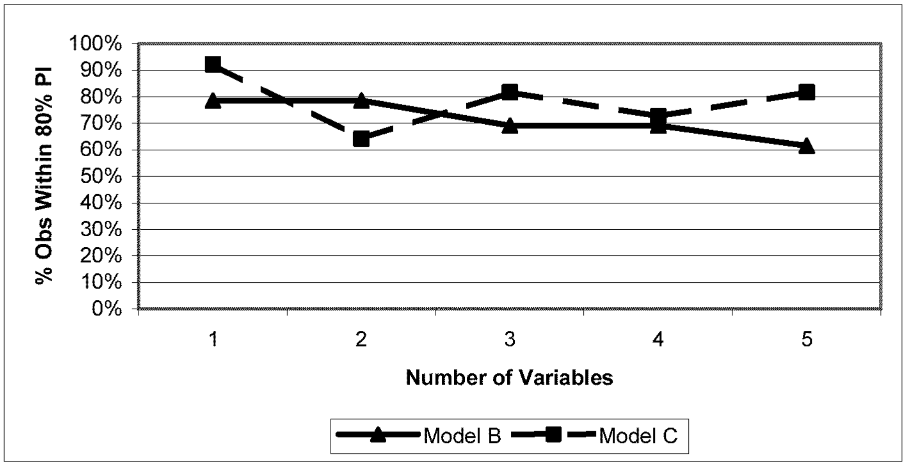


Figure 13. Comparison of PIs for Models B and C

Multiple Regression Results – Model D

We develop Model D to investigate the consequences of not recognizing the mixture distribution of *Engineering %* (continuous and discrete) and overlooking the theories underlying OLS regression, which require a reasonable assumption of both normality and homoskedasticity in the residuals. Model D uses all 90 data points to develop a one-step approach to determining the amount of cost growth that a program will incur. As such, the model produces both negative and positive values for expected cost growth.

Tables 21 and 22 list the best models that we discover through stepwise and OLS regression for each predictor level. In our regressions we find not even one version of Model D that meets the normality and homoskedasticity requirements. Neither do we find a model without influential outliers or with influential outliers that we can remedy through extraction from the data set. We attempt several transformations of the model in order to make the variance constant, but all attempts fail such that the resulting transformation fares worse in a Breusch-Pagan test than the original equation. In other words, Tables 21 and 22 contain nine models which have no statistical grounding, but for which we evaluate their ability to predict cost growth from a pragmatic perspective.

Table 21. Evaluation Measures for Model D

Evaluation Measures	Number of Predictors								
	1	2	3	4	5	6	7	8	9
R^2 Adj	0.1760	0.2266	0.3139	0.3543	0.3280	0.3931	0.3932	0.3869	0.4051
Number of Data Points	90	90	81	81	70	70	70	70	70

Table 22. *P*-Values of Predictor Variables for Model D

Predictor Variables	Number of Predictors								
	1	2	3	4	5	6	7	8	9
Funding Yrs Prod Completed	0.0001	0.0001	0.0001	0.0001		0.0001			
No Maj Def KTR		0.0109	0.0026	0.0054		0.0091	0.0181	0.0260	0.0176
Total Quantity			0.0088	0.0162	0.0161	0.0010	0.0003	0.0225	0.0329
Risk Mitigation?				0.0183	0.0236			0.0260	0.0087
Actual Length of EMD MSIII-MSII in mos					0.0330	0.0310	0.0098		
Maturity (Funding Yrs Complete)					0.0008				
# Product Variants in this SAR					0.0082			0.0269	0.0141
Munition						0.0157	0.0684		
Class at Least S						0.0374			
Funding Yrs R&D Completed							0.0001		
Versions Previous to SAR							0.0472		
Helo							0.0440		
MSIII-based Maturity of EMD%								0.0002	0.0002
N Involvement?								0.0348	0.0235
Class C								0.0129	0.0094
Class U								0.0891	0.0487
Litton									0.0955

From Table 21, we see the general trends of the overall increase in adjusted R^2 with the addition of predictors and the general decrease in the number of data points from 90 to 70 as we increase the number of predictors. Though not listed in this chart, all *p*-values for ANOVA have consistency at the value of 0.0001. We caveat this assessment with the reminder that we view these measures as dubious because of the lack of a theoretical foundation for these models resulting from their failing of normality and constant variance tests.

In Table 23, we see more of a variety of predictor variables than in A, B, or C. In part, this diversity stems from the fact that for Model D we attempt up to a nine-predictor model. However, comparing the range of variables attempted for Models D.1 through D.7 with A.1 through A.7 reveals that Model D uses a third more predictor variables for this range of models than does Model A. We also note that through Model D.6, all predictors have significant *p*-values. The last three models have at least one statistically insignificant predictor at an alpha level of 0.05. These *p*-values have the same problem as the other inferential statistical measures the model generates: without constancy of variance and normally distributed residuals, these *p*-values are potentially erroneous. To what degree they are erroneous, we cannot know. Table 23 shows incremental changes in the model evaluation measures.

Table 23. Incremental Changes in Evaluation Measures for Model D

Evaluation Measures	Number of Predictors								
	1	2	3	4	5	6	7	8	9
Incremental increase in R^2 Adj with additional predictor	0.1760	0.0507	0.0873	0.0403	-0.0263	0.0651	0.0002	-0.0063	0.0182
Ratio of data points to number of variables	90.0	45.0	27.0	20.3	14.0	11.7	10.0	8.8	7.8

We first evaluate how much of the variance of the response variable (*Engineering %*) the model explains through an investigation of the adjusted R^2 measure. In general, we see the adjusted R^2 increase as the number of predictors increases. Figure 14 shows the incremental trend more clearly. As the figure shows, the first variable adds the most to adjusted R^2, while the five and eight-variable models actually decrease adjusted R^2. This information combined with the fact that the seven and nine-variable models do not add much explanation of the variance, leads us to consider either Models D.4 or D.6 as the best models in terms of predictive efficiency alone. We explore the potential significance of the predictors to help us differentiate between the models.

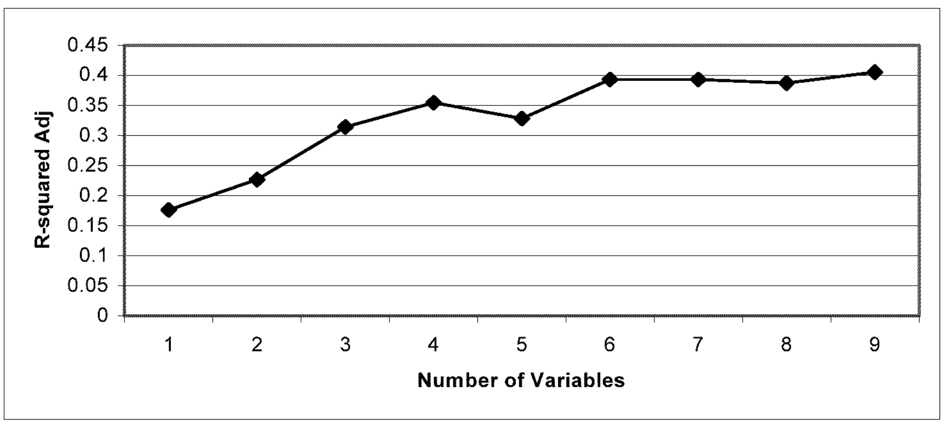


Figure 14. Changes in Adjusted R^2 for Model D

Figure 15 shows how the *p*-values for the least significant predictor and the average *p*-values change with the addition of predictors to the model. The horizontal line marks the upper limit for a significant model at the 0.05 alpha level. Both Models D.4 and D.6 meet this criterion of significance. We note that the least significant *p*-value in Model D.6 is double that of Model D.4. However, the average *p*-value of D.6 remains fairly low (even though one other *p*-value breaches 0.03). The fact that the *p*-value of

D.6 meets the 0.05 criterion, combined with the fact Model D.4 explains only 35 percent of the variance tempts us to accept D.6 over D.4 in order to boost adjusted R^2 to 0.39. However, Table 23 shows that going from Model D.4 to D.6 involves dropping the predictor to data point ratio almost in half from 20.3 to 11.7. We view this drastic decrease along with the modest increase in adjusted R^2 after the addition of the sixth variable as evidence of over-fitting. Thus, discounting the uncertainty associated with the evaluation measures resulting from the unmet model assumptions, D.4 proves the most efficient predictor in the family of the D models.

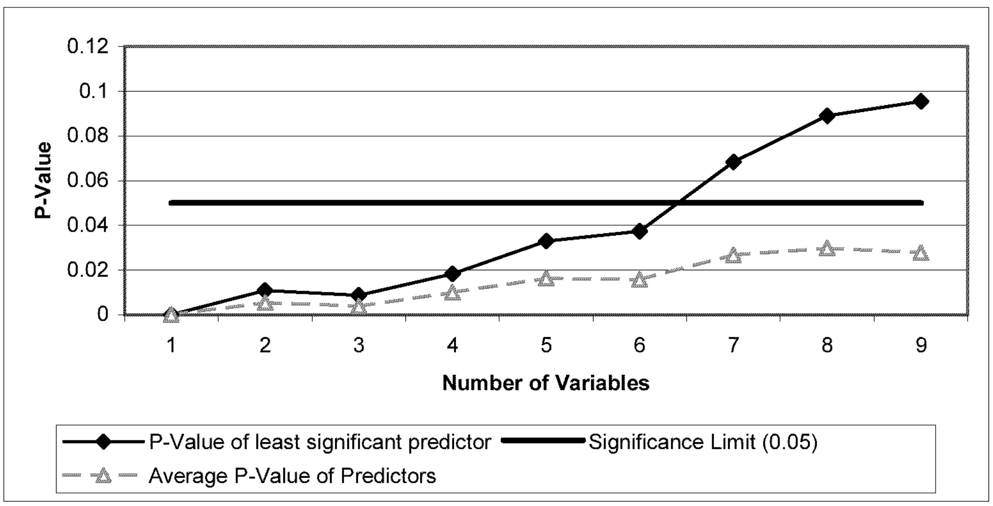


Figure 15. Evaluation of Predictor Variable *P*-Values at a Significance of 0.05

For validation of the data, we perform the same procedures as with Models B and C, and we use the same 25 data points. Unlike Models B and C, this model does not buttress itself upon the assumption that a program will have cost growth. Thus, we do not exclude the 11 validation data points that do not have cost growth. In Table 24 we summarize the results of the validation.

Table 24. Validation Measures for Model D

Validation Measures	Number of Variables								
	1	2	3	4	5	6	7	8	9
Obs Within 80% PI	92.00%	92.00%	86.96%	86.96%	77.78%	77.78%	72.22%	72.22%	72.22%
Avg Width of PI (Eng %)	79.99%	78.01%	79.72%	78.22%	79.95%	76.38%	76.80%	77.33%	78.36%
Obs Below 90% UB	92.00%	92.00%	86.96%	86.96%	77.78%	83.33%	83.33%	77.78%	77.78%
Obs Above 90% LB	100.00%	100.00%	100.00%	100.00%	100.00%	94.44%	88.89%	94.44%	94.44%
Mean Abs Deviation	16.43%	16.93%	19.37%	19.32%	20.09%	19.94%	20.49%	20.97%	20.44%

The results of the range estimation (first four validation measures) show a decrease in ability as the model grows. This is in keeping with the decreasing ratio of data points to variables. The chart shows that the PIs perform adequately, with the LBs capturing all the validation points a majority of the time. The chart also shows that the average width of the PIs remain fairly constant at a width comparable to Models A, B, and C. The mean absolute deviation varies from 16.43 to 20.97 percent. From the measures in Table 24, we do not see enough difference between the models to overturn the previous illation of Model D.4 as the best of the nine models.

Discussion of Models A, B, C, and D

As we point out previously, Models A and B represent the results of obeying the rules of inferential statistics in compiling cost growth models, while Models C and D serve as examples of what happens when we overlook these rules by blindly applying standard regression techniques. Earlier in this chapter we compare Model B with Model C to show that Model B outperforms Model C as a predictive model in both point and range estimation. We now compare Model D as a single-step predictive tool with the two-step approach of using Models A and B to predict whether cost growth will occur and to what degree it will occur. Such a comparison proves difficult and inexact, because the models differ in their methodologies as well as their measures of accuracy, yet we attempt as objective an approach as possible.

Model A produces only binary outcomes, '0' or '1'. One can think of Model D in a similar manner: if Model D predicts a point estimate of zero or less, then we say that Model D predicts a program to have no cost growth. We use Models D.4 and A.7 for the comparison. When we compare the results of the validation using this normalization of Model D's output, we find that Model D's prediction abilities compare well with Model A's on the whole (Table 25). On average, Model A correctly predicts 66.06 percent of the validation points, while Model D correctly predicts 62.87 percent.

Table 25. Percent of Validation Points Correctly Identified as Having or Not Having Cost Growth

% Validation Data	Number of Variables						
	1	2	3	4	5	6	7
Correctly Predicted by A	60.00%	64.00%	61.10%	72.22%	66.67%	69.23%	69.23%
Correctly Predicted by D	68.00%	72.00%	65.22%	60.87%	50.00%	61.11%	

Table 25 seems to indicate that the failure of the normality and constancy of variance assumptions have little effect on the usefulness of the model. Model D.4 proves itself not far inferior to Model A.7 in predicting cost growth. Because of the foundational differences in the model measures, we find this validation procedure the only reasonable quantitative comparison of the two models.

When we consider the performance of Model D versus the performance of Model B at point estimation accuracy, Model D's results do not compete as well. We find first that Model B produces higher adjusted R^2 values than Model D as we show in Figure 16. Model B yields more predictive ability for the number of variables, and none of Model D's versions can compare to the versions of Model B above two predictor variables. Specifically looking at the results of the best model for B, B.3, and the best model for D, D.4, we find that B.3 has an adjusted R^2 value 0.11 higher than D.4, representing an increase in relative predictive power of 31 percent over D.4.

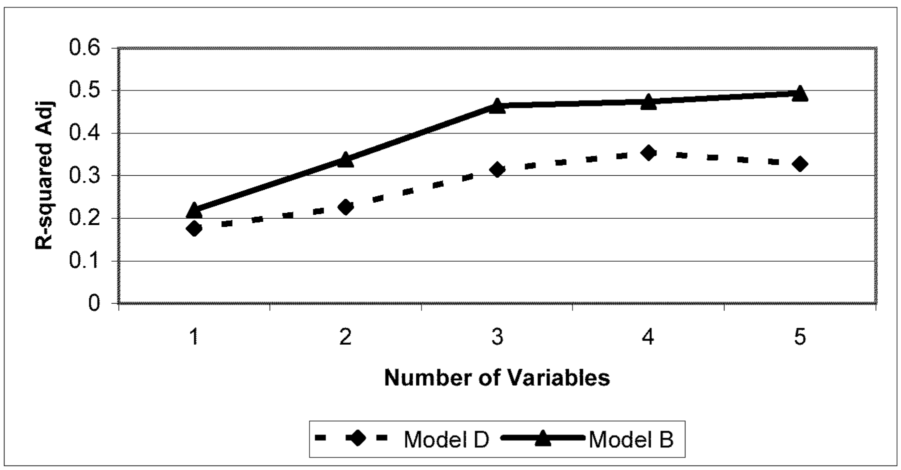


Figure 16. Comparison of Adjusted R^2 for Models B and D

Mean absolute deviation for both models paints a slightly different picture (Figure 17). From these results, we see little distinction between the models. Mean absolute deviation for Model B.3 equals 18.24 percent, smaller that of D.4 which equals 19.32 percent. Thus, on average, the deviations of the actual values from the predicted values of B.3 are about one percent less than D.4's. These results support the results from the adjusted R^2 comparison, concluding B.3 as the better model functionally.

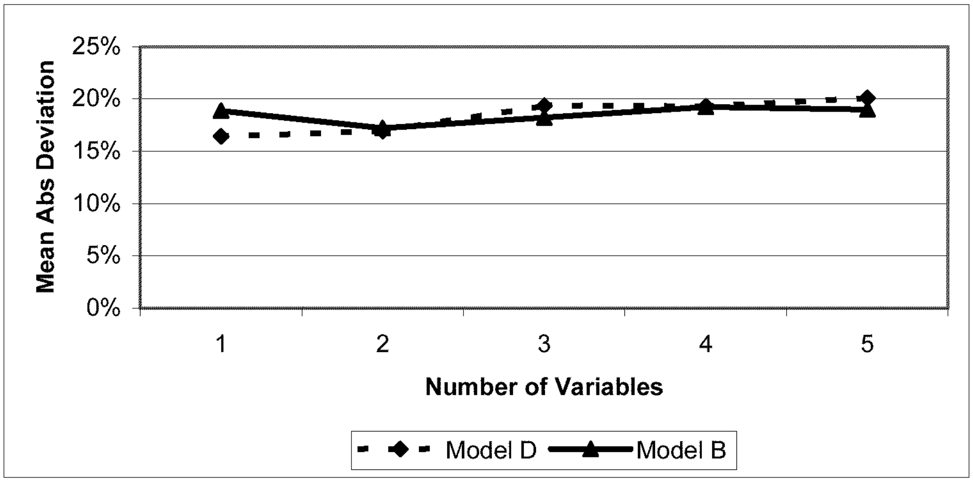


Figure 17. Comparison of Mean Absolute Deviation for Models B and D

In terms of interval estimation (Table 16 and Table 24), Model B outperforms Model D at predicting UBs, while the opposite is true of LBs. Overall, Model D captures a higher percentage of the validation data points within its intervals than does Model B. From these measures, Model D appears to compare very well with Model B in measures of interval estimation. However, one must temper these results first with the fact that Model D uses all 25 validation points, while Model B only uses about half of those data points. These additional data points include negative and zero values. Thus, the comparison between the two model validations has integrity problems.

Following this, Model D predicts intervals that include negative and zero values while at the same time including positive numbers; this jumble of positive, negative, and zero values makes the results difficult to interpret. Finally, and most important, Model D does not actually produce 80 percent PIs. OLS regression produces PIs based on the assumption that residuals have a normal distribution with a constant variance. Since Model D satisfies neither of these requisites, we have no idea what percentage to assign to these intervals. Therefore, Model D appears to perform well, but because we have no idea what exactly it performs, we do not recommend its use. In sum, one might think of Model D like a three-legged dog: it is not put together quite right, but it can be useful.

Discussion of Variables

Until now, this chapter focuses on the model-building and selection process. Now we turn our attention to the variables we use to build these models. Table 26 summarizes the variables used in each of the models described earlier in this chapter. This chart lists overall average significance of each predictor used in Models A through D for all levels of the predictors. The chart also includes the number of times the models use each predictor. We create Figure 18 and Figure 19 to portray how the average significance level and how the frequency of use change from predictor to predictor. These images give an understanding of the number of different predictors that show significant results in predicting engineering cost growth and how often the models use each variable.

Table 26. Significance and Frequency of Predictors for Models A, B, C, and D

Predictor	Mean Significance
Funding Yrs of R&D Completed	0.0004
Funding Yrs Prod Completed	0.0005
Maturity (Funding Yrs Complete)	0.0008
Total Funding Yr Maturity %	0.0024
IOC-Based Maturity of EMD %	0.0024
R&D Funding Yr Maturity %	0.0029
RAND Modification	0.0043
Maturity from MSII (in mos)	0.0052
Length of R&D in Funding Yrs	0.0096
MSIII-based Maturity of EMD %	0.0110
No Maj Def KTR	0.0110
Class C	0.0112
Land Vehicle	0.0132
Air	0.0136
Total Quantity	0.0140
PAUC	0.0161
# Product Variants in this SAR	0.0164
Actual Length of EMD (MSIII-MSII in mos)	0.0176
Length of Prod in Funding Yrs	0.0183
Risk Mitigation?	0.0192
Actual Length of EMD (using IOC-MSII in mos)	0.0244
Svs>1	0.0273
N Involvement?	0.0292
Class At least S	0.0397
Munition	0.0421
Helo	0.0440
Versions Previous to SAR	0.0510
Class U	0.0689
Land	0.0946
Litton	0.0955

Predictor	Frequency (Out of 26 Models)
No Maj Def KTR	15
Actual Length of EMD (MSIII-MSII in mos)	10
Total Quantity	7
MSIII-based Maturity of EMD %	7
Maturity from MSII (in mos)	7
Funding Yrs Prod Completed	7
Length of R&D in Funding Yrs	5
RAND Modification	5
Risk Mitigation?	4
Class At least S	3
Length of Prod in Funding Yrs	3
# Product Variants in this SAR	3
PAUC	3
Class U	2
Versions Previous to SAR	2
Munition	2
N Involvement?	2
Actual Length of EMD (using IOC-MSII in mos)	2
Class C	2
Funding Yrs of R&D Completed	2
Litton	1
Land	1
Helo	1
Svs>1	1
Air	1
Land Vehicle	1
R&D Funding Yr Maturity %	1
IOC-Based Maturity of EMD %	1
Total Funding Yr Maturity %	1
Maturity (Funding Yrs Complete)	1

From Figure 18, we see certain break points where the mean significance measure increases relatively abruptly. The first of these break points occurs after the third predictor. Those variables before this break point we consider as the most significant predictors. We find that all of these variables represent a schedule measure in terms of funding years completed. Between the eighth and ninth variables we see another break. Within the first eight variables, only the indicator variable for modification programs is not schedule related. Thus, we see that schedule criteria dominate the prediction of cost growth. The chart also clearly shows that some of the predictors we use in the models do not have a mean significance at the alpha level of 0.05; these variables have a borderline ability to predict cost growth at best.

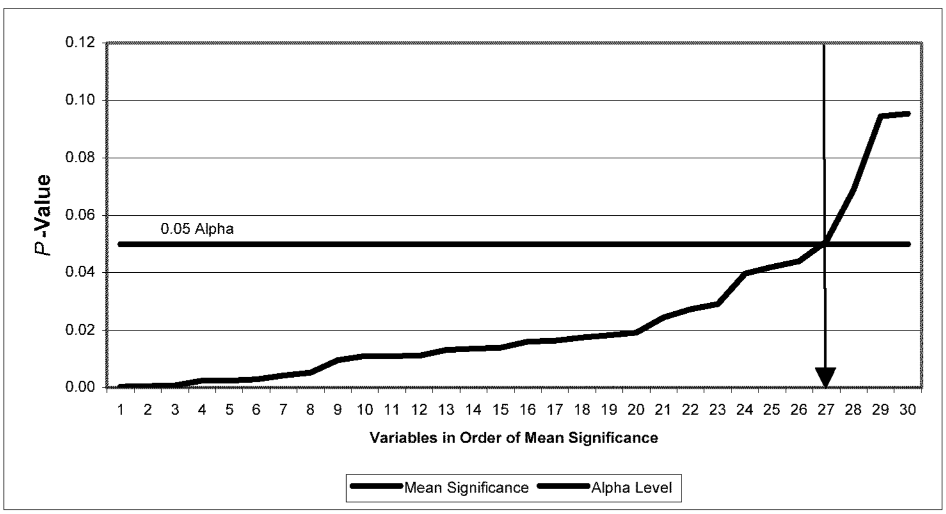


Figure 18. Mean Significance of Predictors for Models A, B, C, and D

Figure 19 displays the frequency of the predictors in the models. From the graph, one can see that a third of the variables occur only once, another third occur two to three times, and a final third occurs from four to fifteen times in the models. Looking at the most frequent third of the variables, schedule variables again appear quite often in the models. Other than schedule variables, the modification identifier, total quantity, whether a major defense contractor worked on the program, and whether engineering risk mitigation existed in the program all occur often in the models. Information from Table 26 and its accompanying figures adverts the reader to variables that tend to predict cost growth best. However, we suggest caution in arriving at conclusions from this data, since the data contains predictors from the statistically questionable Models C and D. A focused look at Models A and B yields more colorable results.

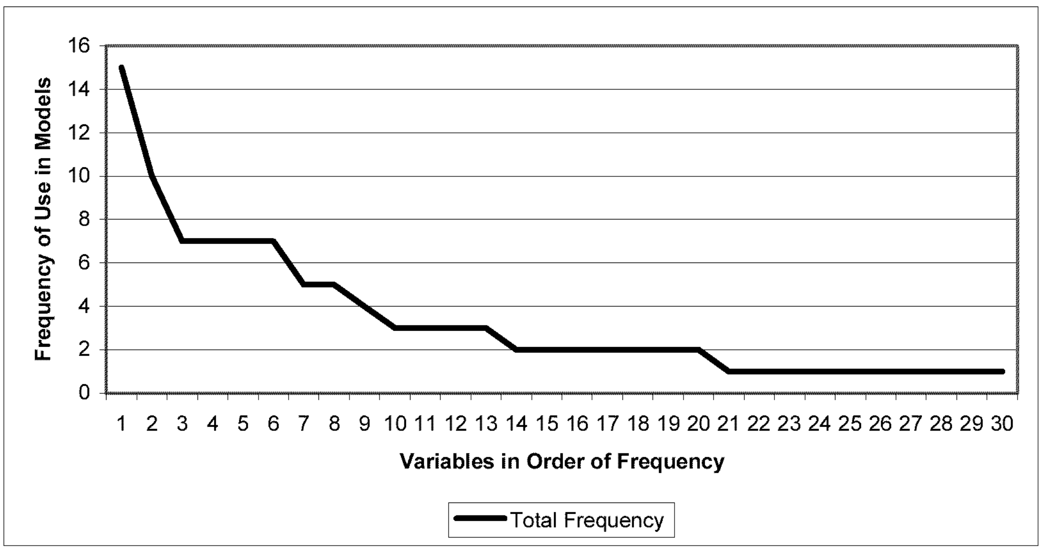


Figure 19. Frequency of Predictors for Models A, B, C, and D

Table 27 displays the predictors of Models A and B. As in the previous table, schedule variables dominate as the most significant and frequent variables used. Within these schedule variables we also find that the modification variable and whether a major defense contractor participated in the program demonstrate relatively high frequency and significance as predictors of cost growth. Service management variables, variables describing physical characteristics and domain of operation, and concurrency variables do not appear on the list of variables used.

Table 27. Significance and Frequency of Predictors for Models A and B

Predictor	Mean Significance
Funding Yrs of R&D Completed	0.0006
Total Funding Yr Maturity %	0.0024
R&D Funding Yr Maturity %	0.0029
Maturity from MSII (in mos)	0.0031
No Maj Def KTR	0.0036
RAND Modification	0.0043
Actual Length of EMD (MSIII-MSII in mos)	0.0063
Length of R&D in Funding Yrs	0.0096
Land Vehicle	0.0132
PAUC	0.0161
Length of Prod in Funding Yrs	0.0183
MSIII-based Maturity of EMD %	0.0189
Actual Length of EMD (using IOC-MSII in mos)	0.0244
Svs>1	0.0273
Class At least S	0.0355

Predictor	Frequency (Out of 12 Models)
Maturity from MSII (in mos)	6
Length of R&D in Funding Yrs	5
Actual Length of EMD (MSIII-MSII in mos)	5
RAND Modification	5
No Maj Def KTR	4
MSIII-based Maturity of EMD %	4
PAUC	3
Length of Prod in Funding Yrs	3
Actual Length of EMD (using IOC-MSII in mos)	2
Class At least S	1
Svs>1	1
R&D Funding Yr Maturity %	1
Total Funding Yr Maturity %	1
Land Vehicle	1
Funding Yrs of R&D Completed	1

In order to compose a single list of ranked predictors based on both mean significance and frequency, we develop a measure that weights the mean significance of the variables by the frequency of use. We call the measure *Overall Importance* (OI), and we calculate it by dividing the significance of a predictor by its frequency. This equation simplifies to the sum of the significance values divided by the frequency squared. The resulting number has no meaning outside its ability to stratify the data according to significance weighted by frequency. Table 28 displays the results of the predictors ranked by OI.

Table 28. Overall Importance of Predictors Models A and B

Predictor	OI
Maturity from MSII (in mos)	0.0005
Funding Yrs of R&D Completed	0.0006
RAND Modification	0.0009
No Maj Def KTR	0.0009
Actual Length of EMD (MSIII-MSII in mos)	0.0013
Length of R&D in Funding Yrs	0.0019
Total Funding Yr Maturity %	0.0024
R&D Funding Yr Maturity %	0.0029
MSIII-based Maturity of EMD %	0.0047
PAUC	0.0054
Length of Prod in Funding Yrs	0.0061
Actual Length of EMD (using IOC-MSII in mos)	0.0122
Land Vehicle	0.0132
Svs>1	0.0273
Class At least S	0.0355

Here again, we see the importance of the schedule variables. In particular, those schedule variables that together sketch an image of where a program exists in RDT&E and, in particular, EMD have the highest OI rankings. Again, the modification and the major defense contractor identifier variables rank high on the list – three and four respectively. Of the four other non-schedule variables, three fall at the end of the list. We insert Figure 20 to give a perspective of the spread of the OI values for these variables.

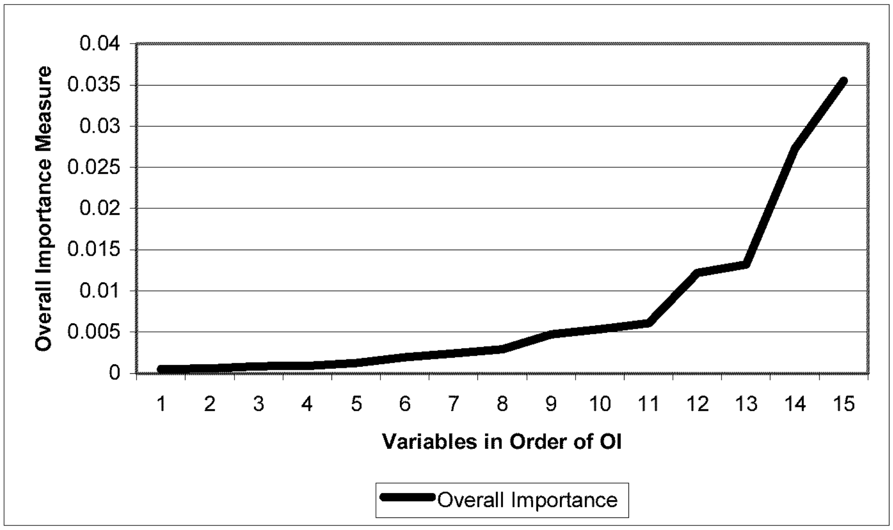


Figure 20. Overall Importance of Predictors Models A and B

The OI graph shows an overall exponential pattern, where the increasing OI values indicate decreasing relative importance of the variables in the models. From this graph, we see the first four variables have almost indistinguishably low OI values. Variables five through eleven gently decrease in importance, and variables twelve through fifteen escalate in OI value (and down in relative importance) very quickly. From these analyses of the model predictors, we gain an understanding of not only the relative importance, but also the magnitude of the stratification of relative importance of the various predictors for cost growth.

Chapter Summary

We analyze four families of models in this chapter, each with several generations of sub models that differ in the number of variables used and the particular variables used. From these subsets we select the best models for each number of variables and compare them using statistical measures of accuracy and significance until we arrive at one best model for each family. We judge Models A.7, B.3, C.4, and D.4 as the best models for each family of model, and then we compare these models with each other. Our study reveals that A.7 and B.3 perform well in determining whether a program will have cost growth and how much cost growth a program will have, respectively. We include the computational forms of these models in Appendix C. C.4 and D.4 seem to perform well, but their lack of conformity with underlying regression assumptions renders the user incapable of accurately interpreting their results.

V. Conclusions

Explanation of the Problem

Cost growth plagues major weapon systems in DoD. The cost estimator's approach to handling cost growth involves increasing cost estimates with cost risk factors to accommodate expected cost growth. Current means of estimating cost risk factors ranges from quantifying expert opinions to developing cost estimating relationships (CERs) from historical data. Reasonable people would agree that the best estimates of cost growth in general come from relationships developed from recent, relevant, and accurate, historical databases. Thus, we seek in this thesis to discover such relationships from such a historical database using regression techniques. We use an approach not found in our literary search: we disaggregate cost growth into separate components in order to seek separate predictor variables for each part. Because this method entails separate analysis for each of the seven SAR-defined constituents of cost growth, this approach has the potential to give more insight into the relationships of variables that might predict cost growth than past research. It also creates the opportunity to build more accurate models.

Limitations

Though we separate cost growth into its components, this study only addresses one of the seven components of cost growth – engineering cost growth. In addition, we only address cost growth in RDT&E dollars and only in the EMD phase of acquisition. Finally, the resulting equations only apply within the range of data used to build them.

Extrapolation beyond these bounds may produce nonsensical results, thus we advise caution with such a use of the models.

Summary of Literature Review Results

A thorough study of recent literature pertaining to cost growth in major defense acquisition systems supports the research of this document. Among the sources we peruse, eleven studies serve to focus us on certain independent variables as candidate predictor variables for cost growth. The scope of these studies differs from the scope of our study in that none of these studies focus specifically on cost growth in the RDT&E budget for the EMD phase. Given this difference in scope between this study and past studies, we consider the applicability of the results of those past studies with an appropriate degree of discretion. We develop from these studies a list of 78 candidate predictor variables for use in this study.

Review of Methodologies

We extract our data from the SARs. In order to have a broad base of programs for measuring the ability of various programs and to include the most accurate data without going too far back in time, we gather all programs from all services that have EMD SAR's recorded for the period 1990 through 2000. We convert all dollar amounts into a common base year, and perform mathematical operations to arrive at predictor variables. We compute our response variable, which we call *Engineering %* for programs. This variable represents the total engineering cost variance in RDT&E dollars divided by the total baseline cost of a program in RDT&E dollars. We convert amounts to base year 2000 dollars for all calculations.

Once we create the database, our exploratory analysis reveals that the response variable has a mixed distribution: a discrete mass representing a large proportion of the data rests at the value zero, while the rest of the data has a continuous distribution. This leads us to the conclusion that we must develop two models for predicting cost growth. The first model, Model A, uses logistic regression to discriminate between those programs that show cost growth and those that do not (grouped with the latter are those programs that experience a negative cost variance). Given that a program experiences cost growth, the second model, Model B, uses multiple regression to determine how much cost growth will occur. At the start of our development of Model B, we find that the response variable of those programs with cost growth has a lognormal distribution. Thus, we transform the response variable via the natural log, and call this Model B. In order to have a baseline to compare the natural log transformation with, we attempt the regression without the transform and label this regression, Model C.

As a potential competitor with the two-step process of Models A and B, we develop a third model, Model D. This model, because of the unusual distribution of the response variable, defies all the assumptions of OLS regression. In fact, we attempt to transform the model, but our attempts only exacerbate the assumption violations. Despite its theoretical shortcomings, we investigate Model D to determine what conclusions, if any, one might reach at its use. For all four models, we set aside approximately 20 percent of the data for validation and use the remaining 80 percent for model building.

Restatement of Results

We find that a seven-variable model provides the best results for the logistic regression. This model accurately predicts about 70 percent of the validation data. The three-variable model provides the best prediction package for Model B. Models C and D produce results that appear similar in effectiveness to Models A and B; however, these models fail the assumptions of normality and constant variance of residuals. We attempt to correct these shortcomings in Model D through several different transformations of the response variable, but find the attempts futile. In addition to the assumption violations, Model D has influential outliers that we cannot remove without creating more influential outliers. Therefore, all models seem to perform well, but only A and B have statistically valid results.

Our results not only establish a case for the applicability of logistic and Y-transformed multiple regression in cost growth analysis, but they also give insight into program characteristics that can be useful to predict engineering cost growth. Overall, the continuous schedule variables provide the most significance and appear more frequently than most other variables. The modification identifier variable and the major contractor identifier variable also perform with statistical significance and frequency that rivals the schedule variables. By identifying predictors of cost growth and their functional relationships to engineering cost growth, we add to contemporary insight into the underlying drivers of engineering cost growth.

Recommendations

Logistic regression provides unique abilities to the cost estimator previously unexplored as far as we can tell from our research. It not only offers the ability to predict whether a program will or will not experience cost growth (50 percent or more = Yes, otherwise No), but also provides the estimator with an estimated probability that a certain program will have cost growth. This allows the user to make predictions more or less conservatively according to a certain percentage assurance desired. In addition to this capability, logistic regression alleviates the estimator from the problematic situation of trying to interpret a linear regression result that indicates the program will experience some negative amount of cost growth within a prediction interval that might include both a negative and a positive estimate.

Moreover, if the distribution of the Model B response variable (*Engineering %*) does not represent an isolated incident, but rather represents a general trait of cost growth databases, then logistic regression proves useful in estimating cost growth. The cost estimating community should consider logistic regression a valid tool and explore its usefulness in other situations where one can translate the response variable into a binary response, rather than rely on OLS where its requisite assumptions will not hold.

In situations where an estimator knows cost growth exists, multiple regression using Model B proves not only theoretically sound, but also demonstrates good point and range estimating capabilities. The cost estimating community should look to this model for estimating engineering cost growth. However, as mentioned earlier, models do not yet exist to estimate the other components of cost growth, and until such time, this model can have utility only in estimating cost growth due to engineering changes.

In Figure 21 we suggest a possible mapping of the seven SAR-defined categories of cost growth to the three AFMC categories of risks included in cost estimates. Given this mapping, this research provides for the foundation of the bulk of engineering risk for RDT&E dollars in the EMD phase. Thus, we pave the way for the potential completion of a historically based model in line with AFMC guidance.

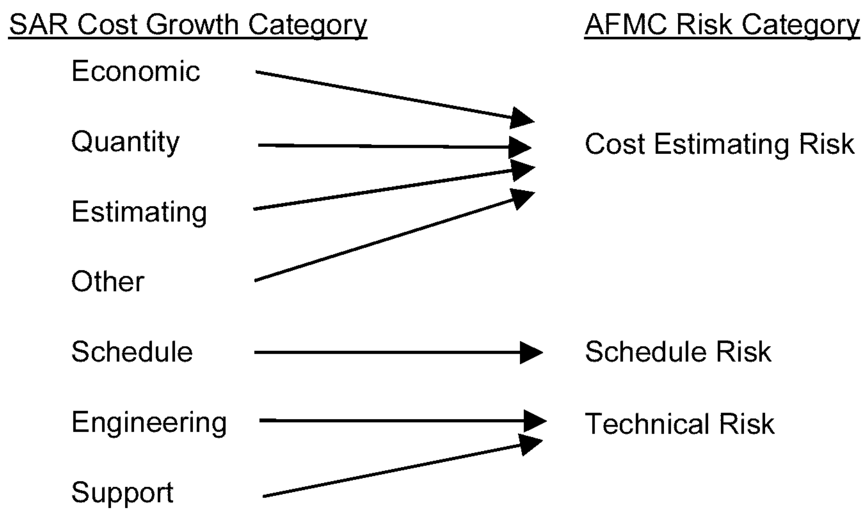


Figure 21. Possible Mapping of SAR Cost Growth to AFMC Risk Categories

Finally, we do not recommend using Models C and D. These models might seem to have some practical ability to estimate cost growth based on their comparable results with the other two models; however, without the underlying assumptions of regression, the interpretations of the results of the models remain ambiguous, and we have no confidence that the process will continue to give similar results over time.

Possible Follow-on Theses:

We encourage the exploitation of the database created during this research for other research topics. We collect a wide range of data in order to develop the dozens of predictor variables explored in this research. Those same data points might prove useful in research of other cost and programmatic areas. Here are some examples:

- Identify programs that did not have significant overruns and evaluate their risk estimating methodology to see if there is a best methodology.

- Accomplish what we did for procurement dollars in the EMD phase.

- Accomplish what we did for the PDRR and procurement phases for both RDT&E and procurement dollars.

- Look for a relationship between overruns and CARD inputs at the time of the DE.

- Analyze the distributions of the overruns across years and fit a curve.

- Look at the autocorrelation of cost growth in each of the four categories of cost growth to see if a relationship exists (this might be along the vane of curve fitting).

- Create a program utilizing the CERs developed from the analysis.

- Experiment with the sensitivity of the models we create to varying inputs.

- Explore the applicability of our results to the Monte Carlo simulation technique of risk analysis.

Appendix A. Seven-Predictor Logistic Regression Model (Model A)

Nominal Logistic Fit for R&D Cost Growth?

Whole Model Test

Model	-LogLikelihood	DF	ChiSquare	Prob>ChiSq
Difference	25.298165	7	50.59633	<.0001
Full	16.778665			
Reduced	42.076830			

RSquare (U)	0.6012
Observations (or Sum Wgts)	61

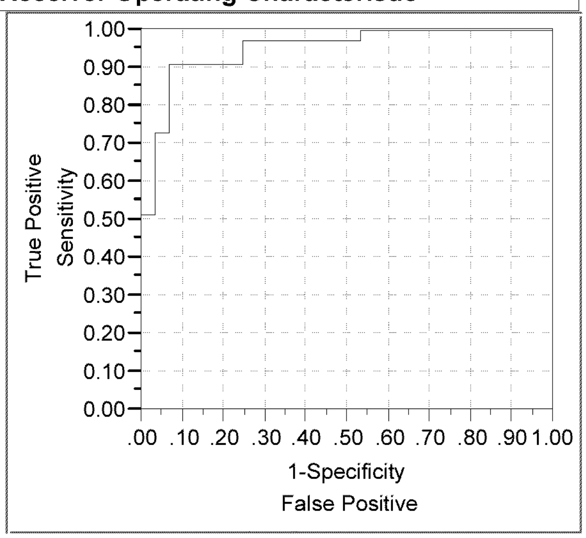
Converged by Gradient

Parameter Estimates

Term	Estimate	Std Error	ChiSquare	Prob>ChiSq
Intercept	9.89273712	3.0519881	10.51	0.0012
Actual Length of EMD (MSIII-MSII in mos	-0.160053	0.0536555	8.90	0.0029
MSIII-based Maturity of EMD %	-2.0396671	0.8368564	5.94	0.0148
RAND Modification?	-4.7892385	1.6482829	8.44	0.0037
Length of R&D in Funding Yrs	-0.5050226	0.1630489	9.59	0.0020
Length of Prod in Funding Yrs	0.49725244	0.153934	10.43	0.0012
Actual Length of EMD using (IOC-MSII in	0.0959051	0.039578	5.87	0.0154
Land Vehicle	-4.8765107	1.9680859	6.14	0.0132

For log odds of 0/1

Receiver Operating Characteristic

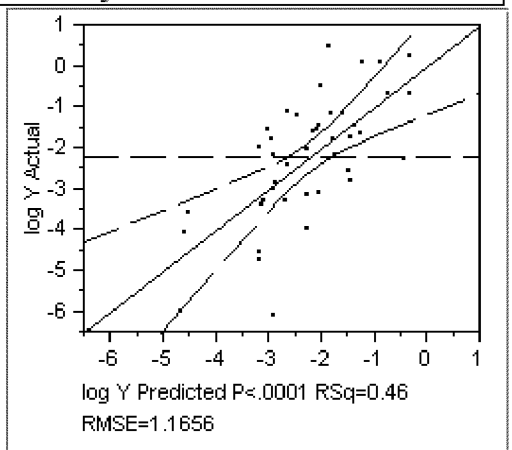


Area Under Curve = 0.94805

Appendix B. Three-Predictor Y-Transformed Multiple Regression Model (Model B)

Whole Model

Actual by Predicted Plot



Summary of Fit

RSquare	0.464491
RSquare Adj	0.422214
Root Mean Square Error	1.165613
Mean of Response	-2.2178
Observations (or Sum Wgts)	42

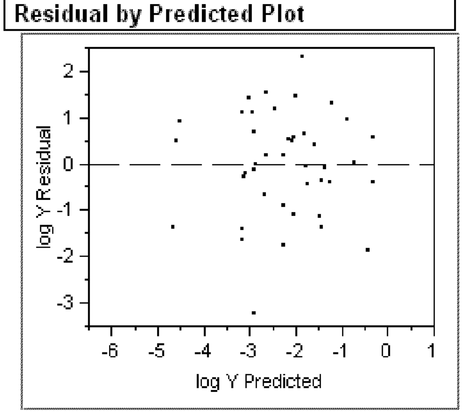
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	3	44.781937	14.9273	10.9868
Error	38	51.628864	1.3587	Prob > F
C. Total	41	96.410800		<.0001

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	-3.38629	0.413	-8.20	<.0001	.
(A1) Maturity from MSII (current calcul	0.0073665	0.002577	2.86	0.0069	1.1107256
No Maj Def KTR	1.3542639	0.415987	3.26	0.0024	1.0340835
Prog Acq Unit Cost	-0.000788	0.000373	-2.12	0.0410	1.0786379

Residual by Predicted Plot



Appendix C. Computational Forms of Models A and B

Computational Form of Model A

Y-Intercept	Coefficients	Input X Values Below:	Predictor Variable
9.892737	-0.160053	174	Actual Length of EMD (MSIII-MSII in mos)
	-2.0396671	95%	MSIII-based Maturity of EMD %
	-4.7892385	0	RAND Modification? (1=Yes, 0=No)
	-0.5050226	26	Length of R&D in Funding Yrs
	0.4972524	18	Length of Prod in Funding Yrs
	0.0959051	234	Actual Length of EMD using (IOC-MSII in mos)
	-4.8765107	0	Land Vehicle (1=Yes, 0=No)

0.836018 Probability of Cost Growth

Computational Form of Model B

Y-Intercept	Coefficients	Input X Values Below:	Predictor Variable
-3.38629	0.0073665	36	Maturity from MSII (in mos)
	1.3542639	0	No Maj Def KTR
	-0.000788	0.22	PAUC (in $M)

0.044101 Estimated Cost Growth $M

Bibliography

Air Force Materiel Command. *AFMC Financial Management Handbook.* Wright-Patterson AFB OH: HQ AFMC, December 2001.

Ayres, Bradley. Class handout, SMGT 643, Systems Acquisition Management. School of Engineering and Management, Air Force Institute of Technology, Wright-Patterson AFB OH. November 2000.

Beck, Charles L. Jr., Nina Lynn Brokaw, and Brian A. Kelmar. *DSMC—A Model for Leading Change: Making Acquisition Reform Work.* Fort Belvoir VA: DSMC Press, December 1997.

Birkler, John, John C. Graser, Mark V. Arena, Cynthia R. Cook, Gordon Lee, Mark Lorell, Giles Smith, Fred Timson, Obaid Younossi, and Jon G. Grossman. *Assessing Competitive Strategies for the Joint Strike Fighter.* Santa Monica CA: RAND, 2001 (MR-1362-OSD).

Christensen, David and Carl Templin. "An Analysis of Management Reserve Budget on Defense Acquisition Contracts," *Acquisition Review Quarterly*, 191-205 (Summer 2000).

Coleman, R. L., J. R. Summerville, M. DuBois, and B. Myers. "Risk in Cost Estimating: General Introduction & The BMDO Approach." Briefing at the 33rd Annual DoD Cost Analysis Symposium. Williamsburg VA. 2 February 2000.

Dameron, M. E., C. L. Pullen, J. R. Summerville, R. L. Coleman, and D. M. Snead. "NAVAIR Cost Growth: Overview of Analysis." Briefing at the Aeronautical Systems Center Industry Cost and Schedule Workshop. Wright-Patterson AFB OH. 24 April 2001.

Department of Defense. *Department of Defense Manual Cost Analysis Guidance and Procedures*. DoD 5000.4-M. Washington: GPO, December 1992.

-----. *Earned Value Management Implementation Guide.* Washington: GPO, 3 October 1997.

Drezner, J. A., J. M. Jarvaise, R. W. Hess, P. G. Hough, and D. Norton. *An Analysis of Weapon System Cost Growth.* Santa Monica CA: RAND, 1993 (MR-291-AF).

Druyun, Darlene. *A Blueprint for Action: Final Report*. Defense Reform 2001 Conference. Washington: 15 February 2001.

Eskew, Henry L. "Aircraft Cost Growth and Development Program Length: Some Augustinian Propositions Revised," *Acquisition Review Quarterly*, 209-220 (Summer 2000).

Garson, David. *PA 765 Statnotes: An Online Textbook.* Textbook for a statistics class taught at North Carolina State University, Raleigh NC. http://www2.chass.ncsu.edu/garson/pa765/statnote.htm. 15 Jan 2002.

Goodman, Clifford S. "TA101: Introduction to Health Care Technology Assessment." A Presentation Prepared for Health Care Professionals by The Lewin Group. Bethesda MD: US National Library of Medicine, http://www.nlm.nih.gov/nichsr/ta101/ta101_c1.htm. 16 January 2002.

Hough, Paul G. *Pitfalls in Calculating Cost Growth from Selected Acquisition Reports*. Santa Monica CA: RAND, 1992 (N-3136-AF).

JMP®. Version 4.0.4, IBM, 51.5M, disk. Computer software. SAS Institute Inc., Cary NC, 2000.

Jarvaise, J. M., J. A. Drezner, D. Norton. *The Defense System Cost Performance Database: Cost Growth Analysis Using Selected Acquisition Reports*. Santa Monica CA: RAND, 1996 (MR-625-OSD).

Knoche, Chris. "Defense Acquisition Deskbook: Selected Acquisition Report." Online Defense Acquisition Reference Guide. http://www.deskbook.osd.mil/default.asp. 31 December 2001.

Merriam-Webster's Collegiate Dictionary. http://www.m-w.com/dictionary. 10 January 2002.

Microsoft® Excel 2000. Version 9.0, IBM, 2.35K, disk. Computer software. Microsoft® Corporation, Redmond WA, 2000.

Neter, John, Michael H. Kutner, Christopher J. Nachtsheim, and William Wassermanl. *Applied Linear Statistical Models*. Boston: McGraw-Hill, 1996.

Obringer, Thomas M. *Analysis of Cost Growth and Cost Composition in the Defense Aerospace Industry*. MS thesis, AFIT/GCA/LSY/88S-7. School of Systems and Logistics, Air Force Institute of Technology (AU), Wright-Patterson AFB OH, September 1988 (AD-A201544).

Office of the Under Secretary of Defense, Department of Defense. *Final Report of the Task Force on Acquisition Improvement*. Memorandum for the Under Secretary of Defense for Research and Engineering. Washington: 23 December 1981.

Singleton, Pamela J. *Estimating Potential Cost Growth of the Most Probable Cost Estimate*. MS thesis, AFIT/GCA/LSQ/91S-11. School of Systems and Logistics, Air Force Institute of Technology (AU), Wright-Patterson AFB OH, September 1991 (AAI-7542).

Terry, Mark F. and Mary M. Vanderburgh. *An Analysis of Estimate at Completion Models Utilizing the Defense Acquisition Executive Summary Database*. MS thesis, AFIT/GCA/LAS/93S-9. School of Systems and Logistics, Air Force Institute of Technology (AU), Wright-Patterson AFB OH, September 1993 (AAK-8756).

Tyson, Karen W., Harmon, Bruce R., and Utech, Daniel M. *Understanding Cost and Schedule Growth in Acquisition Programs.* Ft Belvoir VA: Institute for Defense Analyses Paper P-2967, July 1994 (AD-A284321).

"U.S. Aerospace Cost Risk Analysis Survey," *National Estimator,* 23-32 (Winter 2000).

Westgate, Barbara, from Headquarters Air Force Plans and Programs Branch Directorate of Programs and Blaise Durante, from the Office of the Secretary of the Air Force Management Policy and Program Integration Branch. "Cost Growth Study." Briefing to the Chief of Staff of the Air Force. Headquarters U. S. Air Force, Washington. 21 December 2000.

Whitehead, John. "An Introduction to Logistic Regression." An online tutorial published by a professor in the Department of Economics at East Carolina University, Greenville NC. http://personal.ecu.edu/whiteheadj/data/logit/intro.htm. 15 January 2001.

Wilson, Brian D. *An Analysis of Contract Cost Overruns and Their Impacts*. MS thesis, AFIT/GCA/LSY/92S-8. School of Systems and Logistics, Air Force Institute of Technology (AU), Wright-Patterson AFB OH, September 1992 (AAJ-5471).

Woodward, Peter L. *An Analysis of the Management of Funds for Risk and Uncertainty in the Department of Defense*. MS thesis, AFIT/LSH/LSSR 1-83. School of Systems and Logistics, Air Force Institute of Technology (AU), Wright-Patterson AFB OH, September 1983 (AD-A134409).

Vita

Captain Vincent Sipple graduated from Alief Hastings High School in Houston, TX in 1990. He entered the University of Texas as an undergraduate student in business administration, specializing in accounting. He joined AFROTC in 1992, and was commissioned in 1994.

Capt Sipple's first assignment was at the Human Systems Program Office, Brooks AFB, TX. There he served as a cost analyst and financial manager for a variety of acquisition programs. In 1998, Capt Sipple was assigned to the 8th Comptroller Flight, Kunsan AB, Republic of Korea. There he served as a Financial Services Officer, overseeing the military pay, travel pay, disbursing, and accounting liaison operations for the base. In 1999, Capt Sipple was assigned as the Budget Officer for the 6th Comptroller Squadron, MacDill AFB, FL, where he planned and executed the budget for the 6th Air Refueling Wing. In July 2000, he entered the School of Engineering Management, Air Force Institute of Technology as an Acquisition Management student, specializing in cost analysis. Upon graduation, Capt Sipple will be assigned to the Air Force Cost Analysis Agency in Crystal City, VA.

REPORT DOCUMENTATION PAGE	*Form Approved* *OMB No. 074-0188*

1. REPORT DATE *(DD-MM-YYYY)*	**2. REPORT TYPE**	**3. DATES COVERED** *(From – To)*
09-03-2002	**Master's Thesis**	Jun 2001 – Mar 2002

4. TITLE AND SUBTITLE	
ESTIMATING ENGINEERING COST RISK USING LOGISTIC AND MULTIPLE REGRESSION	**5a. CONTRACT NUMBER**
	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	**5d. PROJECT NUMBER**
Sipple, Vincent, P., Captain, USAF	**5e. TASK NUMBER**
	5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAMES(S) AND ADDRESS(S)	**8. PERFORMING ORGANIZATION REPORT NUMBER**
Air Force Institute of Technology Graduate School of Engineering and Management (AFIT/EN) 2950 P Street, Building 640 WPAFB OH 45433-7765	AFIT/GAQ/ENC/02-02

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)		**10. SPONSOR/MONITOR'S ACRONYM(S)**
ASC/FMCE, AFMC Attn: Mr. Michael J. Seibel 1865 4th St Rm 134 WPAFB OH 45433-7123	DSN: 986-5458 e-mail: Michael.Seibel@wpafb.af.mil	**11. SPONSOR/MONITOR'S REPORT NUMBER(S)**

12. DISTRIBUTION/AVAILABILITY STATEMENT

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

This study explores a two-step procedure for assessing defense acquisition program cost growth using historical data. Specifically, we seek to predict whether a program will experience cost growth and, if applicable, how much costs will increase. We compile programmatic data from the Selected Acquisition Reports (SARs) between 1990 and 2000 for programs from all defense departments. We focus our analysis on cost growth in research and development dollars for the Engineering Manufacturing Development phase of acquisition. We further limit our study to only one of the seven SAR categories of cost growth – engineering cost growth. We explore the use of logistic regression in cost analysis to predict whether cost growth will occur. Using this methodology, we produce a statistically significant model that accurately predicts approximately 70 percent of our validation data. For those programs that have cost growth, we use a multiple regression model (an adjusted R^2 of 0.4645), with a natural log transformation, to predict the expected amount of cost growth. We discover the two-step logistic and multiple regression approach produces desirable results. Finally, we find schedule variables to have the most predictive ability from the 78 candidate independent variables analyzed.

15. SUBJECT TERMS

Logistic Regression, Multiple Regression, Cost Variance, Cost Growth, Selected Acquisition Report, SAR, Cost Growth Statistics, DoD Cost Growth Statistics, DoD Cost Growth Study, DoD Cost Variance Study, Inferential Statistics, Cost Growth in DoD Acquisition Programs, Predicting Cost Growth in DoD Acquisition Programs, Predicting Cost Growth

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
					Edward D. White, USAF (ENC)
a. REPORT	b.ABSTRACT	c. THIS PAGE			**19b. TELEPHONE NUMBER** *(Include area code)*
U	U	U	UU	128	(937) 255-3636, ext 4524; e-mail: Edward.White@afit.edu

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std. Z39-18